

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

Analysis of
Megachirella
fossil plugs gap
in evolutionary
tree of reptiles

PAGE 706

ON THE ORIGIN OF LIZARDS

ECONOMICS

FREE CASH, NO STRINGS

Can universal basic income
help combat poverty?

PAGE 626

EQUALITY

PHYSICS FOR ALL

Boosting minority enrolment
in PhD programmes

PAGE 629

EARTH SCIENCE

THE SHAPE OF WATER

Satellite data track changes
in freshwater availability

PAGE 651

➔ [NATURE.COM/NATURE](https://www.nature.com/nature)

31 May 2018

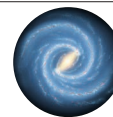
Vol. 557, No. 7707

THIS WEEK

EDITORIALS

CRIME Spanish police introduce algorithm that detects liars **p.612**

PUBLISHING A call to start reproducibility efforts early **p.613**



SUPER-SIZED Disk of stars in the Milky Way is bigger than was thought **p.615**

Get JET set

Confusion over Brexit is adding to the anxiety of staff at a crucial UK research site for fusion energy.

Prime Minister Theresa May conceded on 21 May that a post-Brexit Britain was willing to pay to “fully associate” with Euratom, Europe’s nuclear agency. The details of the arrangement, similar to many that surround the controversial exit of the United Kingdom from the European Union, still have to be ironed out. And among those watching the negotiations with mounting concern are scientists at the Joint European Torus (JET) near Oxford, UK, who currently benefit greatly from Britain’s membership of the agency. The hundreds of researchers at JET receive annual funding of around €60 million (US\$70 million), because Britain is part of Euratom. As it stands, that funding will cease at the end of this year.

The JET facility serves as a key testing ground for ITER, the ambitious experimental fusion reactor being constructed in southern France. For the past three years, JET has been preparing a test run using a mixture of two hydrogen isotopes, deuterium and tritium, to mimic ITER’s planned eventual fuel mix. The test should give the best indication yet of the likely performance of ITER’s particular fusion method — which uses magnetic fields to confine a burning, ionized gas (or plasma) within a doughnut-shaped ring. The run should also help to guide the design of a prototype power plant to follow ITER.

The JET experiment is clearly a crucial project that needs support. But with Brexit looming, where will that support come from?

In theory, the EU can keep paying for JET in the short term. A progress report on the Brexit negotiations, published late last year, says the United Kingdom can continue to pay into and participate in EU funding programmes until December 2020. And Britain has confirmed that it will keep paying its (much smaller) direct share of JET costs until then, too. Moreover, the European Commission (EC) has said that the EU should continue to fund JET; cash for the lab is thought to be included in the EC’s draft programme of fusion-research funding in 2019–20.

But there is a snag. Before the EU can publicly confirm any plans to extend JET’s contract, a number of legislative hoops need to be jumped through. And the process is dragging. The problem lies in how fusion research is funded. A quirk of history means that Euratom’s research funding is allocated in 5-year periods — the current one ending in December 2018 — followed by 2-year top-ups that align the programme with the EU’s 7-year research-funding cycles (the latest of which ends in 2020). Although the top-up is a routine process, it requires the EU Council to approve new legislation, and that has yet to happen.

Renewal of JET’s contract has gone down to the wire before, but the added uncertainty of Brexit is making staff nervous. It hardly helps that the site is repeatedly highlighted in the UK press as a potential casualty of Brexit, rarely with the caveat that its contract should be secure until the end of 2020. JET’s chief executive, Ian Chapman, told *Nature* last year that some top-level staff had already found positions elsewhere. The longer the process drags on, the less attractive JET will seem to researchers.

One wrinkle has already been ironed out: draft text of the EU legislation has been tweaked to allow its fusion programme to include

JET, even if the facility sits outside existing funding schemes. But a vote on the proposed regulation has been delayed by a decision to consult the European Parliament — largely a courtesy that has nothing to do with JET. And because the parliament is unlikely to offer an opinion until September, the final sign-off might now not come until December. No legislation means no research programme, which means no JET contract. The result is that staff at the facility might not officially

know whether they have a job on 1 January 2019 until just days before — let alone be able to do the important deuterium–tritium run.

“Politicians should act to secure JET’s funding.”

The facility itself is ploughing ahead with its preparations for the run, under the assumption that it will be funded for the next

two years. It has no choice but to do so. The planned experiments are key to understanding how plasma will behave in reality, and nowhere else in the world can do the research before ITER is due to begin. Things will probably work out. But the prime minister’s concession regarding Euratom is yet another example of how much her government seems to be making up its Brexit policy as it goes along. Hoping that things will work out is no way to reassure anyone, let alone a basis for strategy. Politicians should act to secure JET’s funding for the next two years — and beyond. ■

Racing hearts

Japan must show that a promising therapy for damaged hearts works as claimed.

As we report in a News story this week (page 619), Japan is set to push ahead with a promising treatment for heart disease that relies on stem cells. It could soon be made available under a fast-track approval system that the country put in place in 2014. Designed to speed access to regenerative therapies, the law allows prospective treatments to be marketed and used as long as they have been proved to be safe. Only a suggestion of efficacy is required — with more-convincing data supposed to be gathered retrospectively from patients who have been given the approved treatment.

The system has its critics — *Nature* among them (see *Nature* 528, 163–164; 2015). The latest move adds further concerns.

The therapy is the work of a physician who was also the first to take advantage of the new law with a related treatment: Osaka University cardiac surgeon Yoshiki Sawa. There is no suggestion that Sawa has not followed the rules, set out by the Pharmaceuticals and Medical Devices Agency. He has. The issue is whether those rules are adequate and appropriate and have the welfare of patients at their heart. They do

not. Treatments of no proven efficacy are being sold to patients (who effectively subsidize the clinical trials to test them). They receive no refund if the therapy is subsequently found not to work. Patients also take risks: they undergo immunosuppression and the surgery itself.

The new study takes induced pluripotent stem cells (iPS cells) that have been banked and characterized to ensure they are safe, and converts them to heart-muscle cells. These are then spread into a thin sheet that is attached to the weakened heart muscle. It is only the second clinical application of iPS cells and is generating excitement around the world. The problem is that the earlier treatment from Sawa — which is ongoing under the fast-track system — has yet to produce convincing results.

In that treatment, approved in September 2015, patients received a sheet of muscle cells made from their own leg tissue, rather than from iPS cells. Called HeartSheet, the muscle sheet is attached to weakened heart muscle that has usually been damaged as a result of a heart attack or plaque build-up and is often the cause of heart failure. The scientists behind the treatment speculate that the muscle cells work by releasing growth factors, not by becoming supporting tissue themselves. Other researchers are sceptical.

Now there are two new treatments being investigated for the same condition, and it's impossible to know yet whether either will work or which might be best for individual patients.

It makes sense that heart-muscle cells (used in the second study) might work better for the heart than leg-muscle cells (used in the first). Indeed, it was reported a decade ago that injecting muscle cells from the leg did not improve heart function (P. Menasché *et al. Circulation* 117, 1189–1200; 2008).

Most physicians hoping to treat heart disease by way of regenerative

medicine have moved on to other strategies, with many looking to heart-muscle cells. That doesn't mean HeartSheet cannot work, but it does raise the question of whether patients who are given it will benefit.

Sawa himself has raised the issue. At a symposium last month touting the new iPS cell trial, he said “leg cells are not good, well, at least not enough”. And the Osaka University web page announcing the iPS cell trial says that HeartSheet was found to be ineffective for more serious cases. Sawa told *Nature* that the cells work in some cases, but

that he expects the new iPS cell therapy to be more effective.

“Treatments of no proven efficacy are being sold to patients.”

All this places a question mark over how the efficacy of HeartSheet can be proved as required. Half way through its scheduled 5-year plan, fewer than 10 patients — of the 60 required by the terms of its approval — have received the treatment. If the trial doesn't make 60, the health ministry told *Nature*, there would either be an extension or the ministry would try to make a decision on the basis of the available data.

Some physicians have called for the HeartSheet tests to end and the data to be assessed before the new iPS cell study can begin. That might be an over-reaction, but pressure on the Japanese government is increasing. The government needs to move quickly to make sure that evaluation of the HeartSheet therapy is as rigorous as promised. As more treatments emerge, officials should make sure that — fast track or not — they have a valid claim to efficacy before being sold to patients.

A therapy for heart disease could be the first iPS-cell clinical breakthrough that Japan so ardently desires. The country shouldn't sell short the promising technology or the patients who hope to benefit from it. ■

False testimony

A lie-detection system being used by Spanish police highlights concerns about algorithms.

If you live in southern Spain, last June was not a good time to lose your smartphone and, as a way of getting an insurance payout, falsely claiming that you had been mugged. Ten police forces in Murcia and Malaga had some extra help in spotting your deceit: a computer tool that analysed statements given to officers about robberies and identified the telltale signs of a lie. According to results published in the journal *Knowledge-Based Systems*, the algorithm was so good at pointing officers towards false claimants that detection of such offences in one week was an impressive 31 and 49 for the respective regions, up from an average of 3 and 12 closed cases over the entire month (L. Quijano-Sánchez *et al. Knowl.-Based Syst.* 149, 155–168; 2018). The government in Madrid is now rolling the system out across the country, and its developers are trying to apply its machine-learning methods to help detect other types of crime.

In this case, the algorithm flagged up suspicious wording (based on a training set of statements known to be true and false), and left it up to the police to question suspects and get them to confess. A person, not a computer, made the final decision. Still, it's another example of the steady march of algorithms and artificial-intelligence (AI) systems into public life and decision-making — and that's a trend that makes some people uncomfortable.

Last week, the UK House of Commons Science and Technology Committee published a report, ‘Algorithms in decision-making’, that summarizes many of those anxieties, and suggests some ways to allay them. It's timely. Also last week, the UK government announced plans to make National Health Service (NHS) data available to companies and others to help build AI-based tools for diagnosing cancer. And the University College London Hospitals NHS Foundation Trust announced a partnership with the Alan Turing Institute, which works

on data science and AI, to find ways of improving health care in the NHS. It aims, for example, to use data sets of previous cases of people who arrive at hospital with abdominal pain, to develop a more effective triage system.

Nature has raised concerns about the development of AI health-care algorithms before, particularly those that seek to diagnose disease (see *Nature* 555, 285; 2018). Although they show great promise, it is crucial that they are developed with proper scrutiny and review of the evidence. That has not always been the case so far.

The UK parliamentary report also discusses a controversial and pertinent issue: how much could and should people who are affected by algorithms' decisions be told about how the software works? This ‘right to explanation’ is included in Europe's new data-protection laws, which came into force last week, although details on how this might change practice are unclear. At present, only France has committed to publishing the code behind algorithms used by the government. More should follow its lead: in evidence to the parliamentary inquiry, the UK government said its departments used such programmes widely; this includes HMRC, the department that calculates and collects tax.

Some witnesses to the inquiry claimed that most people would not understand an explanation of how such software works. Others said that to open the ‘black box’ and lay out how an algorithm works is itself a difficult problem and one compounded by trade secrets. One option, as the report details, is to offer context that helps people to understand the algorithm's workings: to tell someone who has been refused a loan, for example, that the computer helping to make the decision required them to be earning £15,000 (US\$20,000) more a year.

Revealing such details does, of course, allow people to try to game the system. The Spanish police face this problem, too: in describing how their software detects fibs, they are handing advice to those who would lie to them in future about being robbed. This information is already in the public domain, so we're not breaking any confidences by repeating them here: avoid mention of the brand names of what was stolen, don't say the attacker came from behind, and make your statement as long as possible. Still, the Spanish police have an incentive to publicize their system: they hope it will act as a deterrent. In this case, *El Gran Hermano* really is watching you. ■



No reproducibility without preproducibility

Instead of arguing about whether results hold up, let's push to provide enough information for others to repeat the experiments, says Philip Stark.

From time to time over the past few years, I've politely refused requests to referee an article on the grounds that it lacks enough information for me to check the work. This can be a hard thing to explain.

Our lack of a precise vocabulary — in particular the fact that we don't have a word for 'you didn't tell me what you did in sufficient detail for me to check it' — contributes to the crisis of scientific reproducibility. In computational science, 'reproducible' often means that enough information is provided to allow a dedicated reader to repeat the calculations in the paper for herself. In biomedical disciplines, 'reproducible' often means that a different lab, starting the experiment from scratch, would get roughly the same experimental result.

In 1992, philosopher Karl Popper wrote: "Science may be described as the art of systematic oversimplification — the art of discerning what we may with advantage omit." What may be omitted depends on the discipline. Results that generalize to all universes (or perhaps do not even require a universe) are part of mathematics. Results that generalize to our Universe belong to physics. Results that generalize to all life on Earth underpin molecular biology. Results that generalize to all mice are murine biology. And results that hold only for a particular mouse in a particular lab in a particular experiment are arguably not science.

Communicating a scientific result requires enumerating, recording and reporting those things that cannot with advantage be omitted. This harks back to the idea of science as a way to build knowledge through careful experimentation. Ushering in the Enlightenment era in the late seventeenth century, chemist Robert Boyle put forth his controversial idea of a vacuum and tasked himself with providing descriptions of his work sufficient "that the person I addressed them to might, without mistake, and with as little trouble as possible, be able to repeat such unusual experiments".

Much modern scientific communication falls short of this standard. Most papers fail to report many aspects of the experiment and analysis that we may not with advantage omit — things that are crucial to understanding the result and its limitations, and to repeating the work. We have no common language to describe this shortcoming. I've been in conferences where scientists argued about whether work was reproducible, replicable, repeatable, generalizable and other '-bles', and clearly meant quite different things by identical terms. Contradictory meanings across disciplines are deeply entrenched.

The lack of standard terminology means that we do not clearly distinguish between situations in which there is not enough information to attempt repetition, and those in which attempts do not yield substantially the same outcome. To reduce confusion, I propose an intuitive, unambiguous neologism: 'preproducibility'. An experiment

or analysis is preproducible if it has been described in adequate detail for others to undertake it. Preproducibility is a prerequisite for reproducibility, and the idea makes sense across disciplines.

The distinction between a preproducible scientific report and current common practice is like the difference between a partial list of ingredients and a recipe. To bake a good loaf of bread, it isn't enough to know that it contains flour. It isn't even enough to know that it contains flour, water, salt and yeast. The brand of flour might be omitted from the recipe with advantage, as might the day of the week on which the loaf was baked. But the ratio of ingredients, the operations, their timing and the temperature of the oven cannot.

Given preproducibility — a 'scientific recipe' — we can attempt to make a similar loaf of scientific bread. If we follow the recipe but do not get the same result, either the result is sensitive to small details that cannot be controlled, the result is incorrect or the recipe was

not precise enough (things were omitted to disadvantage).

Depending on the discipline, preproducibility might require information about materials (including organisms and their care), instruments and procedures; experimental design; raw data at the instrument level; algorithms used to process the raw data; computational tools used in analyses, including any parameter settings or ad hoc choices; code, processed data and software build environments; or analyses that were tried and abandoned.

Peer review is hamstrung by lack of preproducibility: referees and editors cannot provide serious quality control unless they are given enough information. Preproducibility will bring us closer to the ideals of the Enlightenment, providing crucial evidence about whether a reported result is correct and about how far the result can be generalized.

Science should be 'show me', not 'trust me'; it should be 'help me if you can', not 'catch me if you can'. If I publish an advertisement for my work (that is, a paper long on results but short on methods) and it's wrong, that makes me untrustworthy. If I say: "here's my work" and it's wrong, I might have erred, but at least I am honest. If you and I get different results, preproducibility can help us to identify why — and the answer might be fascinating.

Just as I have pledged not to review papers that are not preproducible, I have also pledged not to submit papers without providing the software I used, and — to the extent permitted by law and ethics — the underlying data. I urge you to do the same. The commitment that Boyle made to the scientific community is even more crucial today. ■

Philip B. Stark is a professor of statistics who specializes in inference at the University of California, Berkeley.
e-mail: stark@stat.berkeley.edu

SCIENCE
SHOULD BE
'SHOW ME',
NOT
'TRUST ME'.

SEVEN DAYS

The news in brief

SPACE

Cool lab launch

NASA's Cold Atom Laboratory arrived at the International Space Station on 24 May. The US\$83-million mission is designed to exploit the microgravity of space to create the coldest point in the known Universe. Once installed, the lab will cool clouds of atoms to a few billionths of a degree above absolute zero so that they form a single quantum state known as a Bose–Einstein condensate. Scientists will use these clouds to probe quantum phenomena in ways that are impossible on Earth. The lab launched on 21 May from the Wallops Flight Facility in Virginia, on a supply rocket.

Instrument fail

The key instrument on the latest next-generation US weather satellite is malfunctioning, the National Oceanic and Atmospheric Administration (NOAA) said on 23 May. The cooling system for the Earth-observing imager did not start up as planned after the March launch of the GOES-17 satellite. The problem affects several infrared and near-infrared detectors, rendering them too warm and degrading the quality of the data that they collect, particularly at night. This will affect the satellite's ability to monitor storms and other weather phenomena. The agency is trying to find a workaround to restore the quality of the information.

Gravity mission

A pair of US–German gravity satellites launched into space on 22 May from Vandenberg Air Force Base in California. The twin spacecraft will monitor how water moves around the planet. The Gravity Recovery and Climate Experiment Follow-On (GRACE-FO) mission picks



XENON

Dark-matter detector draws a blank

The world's largest experiment intended to detect weakly interacting massive particles (WIMPs) has come up empty-handed after collecting data for nearly a year. XENON1T is located 1.4 kilometres underground at the Gran Sasso National Laboratory in central Italy. The experiment looks out for the tiny flashes of light that should be given off when WIMPs — a popular candidate for dark matter, which is thought to make up 85% of the Universe's

matter — collide with atoms in 1,300 kilograms of cold liquid xenon. On 28 May, researchers from the XENON1T collaboration reported at seminars held simultaneously at Gran Sasso and at CERN, Europe's particle-physics laboratory in Geneva, Switzerland, that no such flashes were detected. The data suggest that WIMPs — if they exist — interact even more weakly with ordinary matter than previously thought.

up from the first GRACE spacecraft, which operated between 2002 and 2017 and provided crucial insights into Earth's water cycle and other changes to the planet's mass. The new satellites will measure shifts in surface gravity, which can occur because of processes such as ice loss from polar ice sheets or groundwater extraction for irrigation. The first scientific data from the GRACE-FO mission are expected in about seven months.

EVENTS

Nobel Prize centre

The construction of the Nobel Center's new home in Stockholm has been put on hold after a court opposed

the building's design. The planned 1.2-billion-kronor (US\$137-million) bronze-clad structure is intended to host the annual Nobel prize ceremonies. But Sweden's Land and Environment Court ruled on 22 May that the building would clash with Stockholm's historic waterfront environment. Nobelhuset, the company that runs the centre, says it is waiting to see whether the City of Stockholm, which is developing the project, will appeal against the decision.

North Korea summit

US President Donald Trump has cancelled his planned summit with North Korean leader Kim Jong-un, which had been set for 12 June in Singapore. Trump announced

his decision on 24 May, citing recent remarks by a North Korean official who had described comments by US vice-president Mike Pence as “ignorant and stupid”. US and North Korean officials had been expected to discuss North Korea's efforts to develop nuclear weapons. However, days after the cancellation, media reports suggest that both sides were still preparing for the meeting to go ahead. Also on 24 May, North Korea said that it had dismantled its only known nuclear test site, Punggye-ri. The government invited reporters from several foreign news outlets to watch the demolition but did not allow independent nuclear monitors to attend.

POLICY

The right to try

The US House of Representatives has passed a controversial bill granting some critically ill people the right to access experimental treatments that do not have approval from the US Food and Drug Administration. The bill, voted on by the House on 22 May, had already passed the Senate and will now move on to President Donald Trump, who is expected to sign it into law. Several medical and patient-advocacy groups oppose the 'right to try' legislation, saying that it might offer people false hope and expose them to unnecessary risks.

Bear hunt approved

Wyoming officials approved regulations for the hunting of up to 23 grizzly bears (*Ursus arctos horribilis*, pictured) in the area around Yellowstone National Park. Grizzlies in the park are still off-limits to hunters, as are those in nearby Grand Teton National Park. The decision came on 23 May, less than a year after the US government removed grizzly bears in the greater Yellowstone ecosystem from the endangered-species list. The hunt proposal, put forward by the Wyoming Game and Fish Department in February, stirred controversy over whether this population of bears had recovered from



decades of hunting and habitat loss. In April, 73 scientists asked Wyoming Governor Matt Mead to halt the hunt until independent experts could review the proposal.

Brexit plan emerges

The United Kingdom has called for a close partnership with the European Union on science and innovation after Brexit. A document published on 23 May by the UK government department overseeing negotiations to leave the EU outlines plans for a future science and innovation pact. The proposal includes access to EU-funding programmes and research infrastructure. It also calls for wider agreements on data sharing, intellectual property and the movement of researchers around the EU. The document, which will be used in discussions with EU officials, also makes it clear that the United Kingdom is willing to respect the remit of the Court of Justice of the European Union in relation to

its participation in relevant EU science programmes. Prime Minister Theresa May initially said she did not want the court to have any jurisdiction in the United Kingdom after Brexit.

HEALTH

Nipah vaccine

Two US biotechnology firms will receive millions of dollars to develop a vaccine against the rare but deadly Nipah virus. The World Health Organization lists the infection, transmitted to humans by fruit bats (*Pteropus* spp.), as a priority for research and development. On 24 May, the Coalition for Epidemic Preparedness Innovations, a multimillion-dollar initiative to develop and stockpile vaccines, announced that it would give US\$25 million to Profectus BioSciences and another company, Emergent BioSolutions, which will provide technical and manufacturing support. There is currently no approved

vaccine or treatment for Nipah virus, which kills about 75% of those infected. An outbreak in southern India this month has killed at least ten people.

FACILITIES

Telescope first

A pioneering telescope set-up — the first optical telescope permanently dedicated to tracking the gaze of a radio telescope — launched in South Africa on 25 May. The MeerLICHT instrument (Dutch for 'more light') in Sutherland is linked to South Africa's 64-dish MeerKAT radio telescope near Carnarvon and will observe for the next 5 years. By pointing the telescopes at the same part of sky at the same time, researchers hope to be able to study astronomical events in many wavelengths, which could reveal the causes of enigmatic, short-lived astronomical events such as fast radio bursts. MeerLICHT will focus on such transient events and on detecting possible sources of gravitational waves. The €1-million (US\$1.2-million) project is a collaboration between Dutch, South African and UK institutions.

PEOPLE

New AAS president

Australian biochemist and molecular biologist John Shine is the new president of the Australian Academy of Science (AAS). Shine, a pioneer in human gene cloning, most recently studied the genetics of inherited kidney disorders at the Garvan Institute of Medical Research in Sydney, where he was also executive director from 1990 to 2011. For the past six and a half years, he has served as chair of Australian biotechnology giant CSL, headquartered in Melbourne. Shine started his five-year term at the AAS on 24 May, replacing chemist Andrew Holmes.

NATURE.COM

For daily news updates see:

www.nature.com/news

TREND WATCH

Opinions about which research contributions deserve authorship credit on a paper vary markedly across scientific disciplines — and even within fields. In a survey of nearly 6,000 researchers, most said they would grant authorship for data interpretation or manuscript drafting. But nearly half would almost never or only sometimes do so for people who secured their funding. Social scientists tended to assign less value to proposing ideas than did researchers in the pure, applied and natural sciences.

A QUESTION OF CREDIT

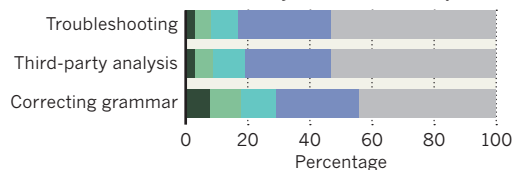
A survey of nearly 6,000 researchers across scientific fields reveals which contributions to research papers tend to attract authorship credit, and which don't.

■ Almost always ■ Usually ■ Often ■ Sometimes ■ Almost never

Contributions that most researchers say deserve authorship credit



Contributions that fewest researchers say deserve authorship credit



NEWS IN FOCUS

PHYSICS From pyramids to nuclear waste, muons help to probe the impenetrable **p.620**

CLEAN ENERGY Testing begins at plant that burns fossil fuels without emitting carbon **p.622**

BIOTECHNOLOGY Universities scramble to protect patents on lucrative antibodies **p.623**



ECONOMICS Experiments test 'universal basic income' schemes **p.626**

THE ASAHI SHIMBUN VIA GETTY IMAGES



At a press conference in Tokyo, cardiac surgeon Yoshiaki Sawa announces plans to use tissue derived from induced pluripotent stem cells to treat heart disease.

CLINICAL RESEARCH

Reprogrammed stem cells approved to mend hearts

Japanese study is only the second application of induced pluripotent stem cells in people.

DAVID CYRANOSKI

Scientists in Japan now have permission to treat people who have heart disease with cells produced by a revolutionary reprogramming technique. The study is only the second clinical application of induced pluripotent stem (iPS) cells. These are created by inducing the cells of body tissues such as skin and blood to revert to an embryonic-like state, from which they can develop into other cell types.

On 16 May, Japan's health ministry gave doctors the green light to take wafer-thin sheets of tissue derived from iPS cells and graft them onto diseased human hearts. The team, led by cardiac surgeon Yoshiaki Sawa at Osaka University, says that the tissue sheets can help to regenerate the organ's muscle when it

becomes damaged, a symptom of heart disease that can be caused by a build-up of plaque or by a heart attack.

"It will excite worldwide attention, as many groups are working in the same direction," says Thomas Eschenhagen, a pharmacologist at the University of Hamburg in Germany and chair of the German Centre for Cardiovascular Research.

The treatment will initially be given to three people over the next year. The team will then seek approval to conduct a clinical trial in around ten patients. If it proves safe, the treatment could then be sold commercially under Japan's fast-track system for regenerative medicine.

The system, introduced in 2014, aims to speed the availability of potentially life-saving

procedures. But critics say the system is flawed because it allows treatments to be on sale to patients before sufficient data have been collected showing that the procedures work.

MENDING BROKEN HEARTS

In their technique, Sawa and his colleagues use iPS cells to create a sheet of 100 million heart-muscle cells. From studies in pigs, the team has shown that grafting these sheets of cells — each 0.1 millimetres thick and 4 centimetres long — onto a heart can improve the organ's function. Sawa says that the cells do not seem to integrate into the heart tissue. He thinks that instead they release growth factors that help to regenerate the damaged muscle.

Scientists say one advantage of the sheets is that they create their own cellular matrix, ►

► and can maintain their structure without the need for scaffolding made from foreign materials, a feature of some other engineered tissues.

"It is a very elegant and clever way to deliver cells," says Philippe Menasché, a cardiac surgeon at the Georges Pompidou European Hospital in Paris, who has also experimented with making tissue sheets.

Pharmacologist Wolfram-Hubertus Zimmermann at the University Medical Centre Göttingen in Germany, who is also developing an iPS treatment for heart disease, says that the latest trial is based on work conducted by Sawa and other colleagues in Japan over the past 15 years.

Once Sawa's team has treated its three patients, it will apply to conduct a clinical trial involving a further seven to ten people. If the treatment proves to be safe, and shows some signs of working, it can be approved for sale under the accelerated system. This allows researchers to bypass expensive large-scale clinical trials aimed at proving efficacy, and instead to use small pilot trials to show that the therapy is safe and shows an indication of efficacy.

But some researchers say the bar for approving therapies for commercial use is too low. Even if the cells are found to be safe, there are risks associated with any surgery, and patients could give up other therapies for a treatment that might not work. Ethicists and regulators say the benefits of any new therapy must outweigh the risks.

Yoshiki Yui, a cardiologist at Kyoto University, says that, as well as meeting the requirements for safety, researchers should show that their treatment is effective, which would require testing it in larger numbers of people than are currently required. The evaluation process should also use randomized, controlled clinical trials, the gold standard for demonstrating efficacy in medical research, he says.

The iPS-cell therapy has potential, Yui adds, but under the current approval system, "we won't know if it works or not" because it won't have been tested in a controlled trial. "The biggest problem is there's no adequate system of evaluation in Japan," he says.

A spokesperson for the health ministry told *Nature* that the current approval system is sufficient because researchers must still show that a treatment works even if it has been approved for commercial use.

Sawa agrees that a control group is important for proving efficacy, but notes that he is abiding by Japan's rules, which don't require this before a treatment is made commercially available.

He says the health ministry's approval is an acknowledgement that the therapy "is scientifically and ethically justified" to be tested in patients. "Whether it really works, [we] have to find out now," he adds. ■ [SEE EDITORIAL P.611](#)

PHYSICS

Muography makes its mark

Little-known particles called muons are helping to map the insides of pyramids, and to spot missing nuclear waste.

BY ELIZABETH GIBNEY

The muon is going mainstream. The particle, a heavy version of the electron that constantly rains down on every square centimetre of Earth, is little known outside particle physics — but last year it helped archaeologists to make a stunning discovery of a previously unknown chamber in Egypt's Great Pyramid.

Volcanologists and nuclear engineers are also finding new uses for muography, which harnesses muons to probe the innards of dense structures. The first companies are looking to cash in.

"The discovery in the pyramids last year has really put muography on the map," says David Mahon, a physicist at the University of Glasgow, UK, who co-organized an international meeting called Cosmic-ray Muography, sponsored by the Royal Society and held on 14–15 May in Newport Pagnell, UK.

MUONS ARE EVERYWHERE

Muons have the same negative charge as electrons but 200 times the mass. They are made when high-energy particles called cosmic rays slam into atoms in Earth's atmosphere. Travelling at close to the speed of light, muons shower Earth from all angles. Every hand-sized area of the planet is hit by roughly one muon per second, and the particles can pass through hundreds of metres of solid material before they are absorbed.

Their omnipresence and penetrating power makes muons perfect for imaging large, dense objects without damaging them, says Cristina Cârloganu, a physicist at the Clermont-Ferrand Physics Laboratory in France. The denser materials are, the more energy they absorb from the particles, so physicists can track how often muons of different energies reach detectors placed around a target, and compare that with the expected rate without an obstacle, to build up a 3D profile of the density of the interior.

Physicists have been experimenting with the technique since the 1950s, including an unsuccessful search for hidden chambers in the second-largest pyramid at Giza (L. W. Alvarez *et al. Science* **167**, 832–839; 1970). But the room-sized detectors were expensive and impractical, says

Raffaello D'Alessandro, a particle physicist at the University of Florence, Italy, and a co-organizer of the muography meeting. They could weigh more than 10 tonnes and relied on muons' ability to ionize particles of sometimes explosive gases.

CHANGE OF TACK

Ways to track the paths of charged particles more precisely — developed at facilities such as CERN, Europe's particle-physics laboratory near Geneva, Switzerland — have made for safer, smaller and more-sensitive muon detectors. They can now be as compact as a few square metres and can run on solar panels, making it possible to take them to remote field sites.

Volcanoes have become a popular target for the technique, thanks to pioneering work by researchers in Japan.

"It's a new, very specialist technique that comes from the high-energy-physics world."

Mapping lava channels, which absorb less energy from muons than does the dense surrounding rock, could one day help to predict eruptions, says Cârloganu.

This year, researchers will try to image the solidified plug of lava inside Italy's Mount Vesuvius.

Smaller devices are also being used in archaeology, says Giulio Saracino, a physicist at the University of Naples Federico II in Italy. He and his team have mapped cavities and tunnels under Mount Echia, a settlement in Naples that has been occupied since the eighth century BC. They also plan to look for a rumoured aqueduct beneath the nearby ancient city of Cumae.

A spate of commercial applications for muography — five were presented at the conference — probe smaller samples, such as drums of nuclear waste. These applications often use a slightly different technique, which tracks how muons change direction when they hit atomic nuclei in a material.

By placing detectors on both sides of a sample, physicists can recreate a particle's trajectory. And because the angle of deflection correlates with the density of the substance the muon hits, studying these paths can help to create a density map of the material being



Muon detectors are now small enough to take to field sites such as the Great Pyramid of Giza in Egypt.

probed. Engineers can use this method to spot stray fragments of uranium inside containers of nuclear waste, even if it is encapsulated in concrete or steel.

“To get information about what is deep in the centre, muons are pretty much the only thing that can do that,” says Mahon. He directs a firm called Lynkeos Technology based in Glasgow, which will start imaging nuclear-waste samples next month at the UK National Nuclear Laboratory at Sellafield.

In the United States, trials at the Los Alamos

National Laboratory in New Mexico have found that similar technology can spot where fuel rods have been removed from casks of spent fuel. Just four stolen fuel rods would provide enough plutonium to build a primitive nuclear weapon, Los Alamos physicist Christopher Morris told the conference.

Israeli firm Lingacom, based in Tel Aviv, is also investigating using the technique in security screening, for example at border crossings, to inspect containers for smuggled nuclear material. Other firms plan to use muography

to track the wear of oil-industry pipelines and search for minerals in old mines.

But in many academic fields, the technology is still greeted with shrugs and quizzical looks. Despite finds such as the Great Pyramid’s hidden chamber, the technology is still relatively unproven. “It’s a new, very specialist technique that comes from the high-energy-physics world,” says Saracino. “The first time I say to geologists that we have muon technology, they say, ‘What are muons?’ They are fascinated, but also a little bit wary.” ■

DIVERSITY

Fewer African American men going into medicine

Diversity advocates seek strategies to correct alarming decrease.

BY GIORGIA GUGLIELMI

Even as US diversity initiatives try to increase the representation of minority ethnic groups in science and medicine, the proportion of black men pursuing such careers is reaching historic lows. In 1986, 57% of African American medical-school graduates were men — but by 2015 that share had dropped to just 35%, even as the total

number of black graduates had increased.

Given the extent of racism and discrimination, “it’s difficult for black males to be able to progress,” says Cato Laurencin, a surgeon-scientist at the University of Connecticut in Farmington. Laurencin chaired a workshop on the issue that was convened last November by the US National Academies of Sciences, Engineering, and Medicine and the Cobb Institute, a non-profit group in Washington DC that studies

health disparities and racism in medicine.

A report from the workshop, released on 18 May, examines factors that contribute to the growing absence of black men in science and medicine, as well as current models and strategies for boosting participation (see go.nature.com/2lo4p3b).

Although more African American students attend medical schools today than 30 years ago, the increase is due to greater numbers of ►

► black women training to be physicians. The proportion of men among African-American medical students decreased by more than 20% over the same period. Data from the Association of American Medical Colleges show that, in 2015, 41% of black male applicants were accepted into medical school — the lowest rate across all genders and ethnicities. “This is a crisis that affects not only blacks, but also our national ability to have excellence in science and medicine,” Laurencin says.

Racial diversity in the medical professions can help to address health inequalities. Studies have shown that people from minority groups receive better care when their physicians have similar backgrounds.

“Having racial diversity leads to not just more doctors, but also better-prepared doctors who go into communities of colour,” says Liliana Garces, an education researcher and legal scholar at the University of Texas at Austin. She adds that one promising strategy for increasing diversity in medical schools is reducing the admission procedure’s emphasis on standardized tests, which “don’t end up capturing the student’s potential, and only contribute to more racial inequities in the student body”.

Ross University School of Medicine in Portsmouth, Dominica, accepts students from under-represented minorities with lower standardized test scores and grade point averages than white applicants. The university — which has campuses in Dominica and the United States — gives these students educational support during the first semesters of medical school and connects them with a mentor from a similar background.

Environments where black men can build a community help to improve graduation rates, Laurencin says. And programmes that give financial support to undergraduate students of colour and provide early exposure to research increase representation in science, technology, engineering and mathematics PhD programmes.

But Freeman Hrabowski, president of the University of Maryland, Baltimore County, which runs one such programme, notes that universities and medical schools need funding to expand these efforts. “Without funding,” he says, “there is no serious commitment.” ■



A demonstration power plant run by NET Power in Houston, Texas.

ENERGY

Zero-emissions plant begins key tests

Start-up firm NET Power is developing a new approach to capturing and storing carbon.

BY JEFF TOLLEFSON

A team of engineers in La Porte, Texas, has spent the past several weeks running tests on a prototype power plant that uses a stream of pure carbon dioxide — not air — to drive a turbine. If the zero-emission technology developed by NET Power in Durham, North Carolina, succeeds, it could help to usher in an era of clean power from fossil fuels.

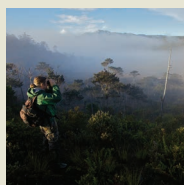
The company broke ground on the roughly

25-megawatt plant in March 2016, after raising US\$140 million for the project. It completed construction last year. It is now running a battery of tests on the combustor that powers the plant, a one-of-a-kind device built by the Japanese industrial giant Toshiba. If the tests go as planned, NET Power will hook up the turbine and begin generating electricity later this year.

Officials say everything is running smoothly so far. “We’re still smiling,” says chemical engineer Rodney Allam, the facility’s lead designer.


**MORE
ONLINE**

TOP NEWS



Indonesian plan to clamp down on foreign scientists draws protest
go.nature.com/2iwxpcn

MORE NEWS

- Hybrid human-chicken embryos illuminate key developmental milestone
go.nature.com/2sfi5in
- Why North Korea’s denuclearization plan doesn’t convince this nuclear expert
go.nature.com/2sehw8Q

NATURE PODCAST



Boosting diversity in physics grad programmes; and life’s recovery after a massive asteroid impact
nature.com/nature/podcast

TIM LAMAN/NATUREP.L.COM

Allam is now a partner with 8 Rivers, a technology company in Durham that co-owns NET Power with Exelon, a major electricity provider in Chicago, Illinois, and McDermott International, an energy-services company in Houston, Texas.

What separates the La Porte facility from a standard power plant is the CO₂ cycle at its core. A conventional power plant burns fossil fuels to generate steam that drives a turbine — and it also emits CO₂ as a by-product.

By contrast, NET Power will drive its turbine with a loop of hot, pressurized CO₂. The first step is to fill the system with CO₂, which must then be heated to drive the turbine — much like a conventional power plant heats water to create steam.

The combustor then ignites a mixture of natural gas and oxygen, which is extracted from the atmosphere — at almost no cost. That gives it an edge over existing technologies for stripping CO₂ out of a conventional power plant's exhaust; these drive up costs while sapping around 20% of the plant's power.

ENERGY ECONOMICS

The result is a stream of pure CO₂ that can be buried or put into a pipeline — rather than the atmosphere — at almost no cost. That gives it an edge over existing technologies for stripping CO₂ out of a conventional power plant's exhaust; these drive up costs while sapping around 20% of the plant's power.

Allam says that, if all goes well, NET Power's technology will produce electricity as cheaply and efficiently as a conventional,

modern gas-fired power plant — and earn extra revenue by other means. For instance, oil companies might buy the plant's excess CO₂ and pump it into their wells to boost oil production. NET Power could also sell nitrogen and argon produced by the plant's air separator.

A coal-fired power plant in Houston that is equipped with a competing CO₂-capture technology is already delivering the gas it collects to a nearby oil field. The \$1-billion Petra Nova project came online in January 2017. It uses an

"If the plant does everything they say, it's hard to imagine why you would want to build a traditional power plant."

amine-based solvent to capture about one-third of the emissions from a single power-generating unit — up to 1.6 million tonnes of CO₂ annually.

But the project — a joint venture between NRG Energy in Princeton, New Jersey, and JX Nippon Oil and Gas Exploration in Tokyo — depended on both a \$190-million grant from the US Department of Energy and additional oilfield revenue to turn a profit, says Daniel Cohan, an atmospheric scientist at Rice University in Houston. By contrast, he notes NET Power's claim that its power plant will turn a profit even before it begins selling CO₂.

"If the plant does everything they say, it's hard to imagine why you would want to build a traditional power plant," Cohan says. "But there are still a lot of ifs ahead."

One major challenge will be ensuring proper combustion of oxygen and methane in the presence of CO₂, which normally acts as a fire extinguisher. NET Power is several months behind schedule on this task, but project officials say that was the result of Toshiba's decision to test the plant's combustor on site rather than sending it to an independent test facility; that meant installing and reconfiguring equipment at the otherwise complete plant.

Once the project begins producing electricity, NET Power engineers must also show that the plant operates as efficiently as advertised, says Howard Herzog, who studies carbon capture and sequestration at the Massachusetts Institute of Technology in Cambridge. The challenge, he says, will be to address the inevitable problems that arise when engineers are building the first-of-a-kind facility without sacrificing energy efficiency or driving up costs.

NET Power officials say they are ready to take advantage of recently expanded US government tax credits for carbon capture and sequestration projects — beginning with a proposed 300-megawatt plant that could be operational by 2021. But the company's chief executive, Bill Brown, says the firm isn't reliant on subsidies, and is already seeking customers and manufacturing partners abroad. It is also looking at potential markets for CO₂, which could soon become a cheap chemical feedstock.

"We don't like to rely on policy around here," Brown says. "We like to rely on science." ■

INTELLECTUAL PROPERTY

Rush to protect billion-dollar antibody patents

A US federal court decision has left biotech working to preserve intellectual-property rights.

BY HEIDI LEDFORD

Universities and biotechnology companies in the United States are scrambling to protect some of their most valuable assets: patents on antibodies. These immune-system molecules form the basis of drugs that rake in about US\$100 billion per year. But securing intellectual-property rights to antibodies has become much more difficult, under guidelines released in February by the US Patent and Trademark Office (USPTO).

The revised rules come after a federal court decision last October narrowed the scope of antibody patents — including those that have

already been handed out. "People are still trying to make sense of it," says Ulrich Storz, a patent attorney at Michalski Hüttermann & Partner in Düsseldorf, Germany. "These were very powerful patents."

Storz and others will discuss the implications of the shift on 6 June as part of a panel at the Biotechnology Innovation Organization annual meeting in Boston, Massachusetts.

BROAD PROTECTIONS

Antibodies are proteins made by the immune system that bind to a specific target, such as a protein produced by a microbe, to interfere with its ability to promote disease.

This has made them powerful drugs against some illnesses.

Therapeutic antibodies are structurally complex, and in many cases, changes to their amino-acid sequences will not affect their function. So a patent based solely on an antibody's sequence might be vulnerable to competition, says Barbara Rigby, a patent attorney at Dehns in Brighton, UK. A competitor could tweak the sequence to create a new antibody that performed the same function.

In addition, for many years researchers lacked the technology to sequence an antibody, to define how it bound its target or to introduce specific changes to its ►

► structure. Given these challenges, the USPTO routinely granted broad patents that covered the suite of antibodies that attached to a particular target, rather than a specific antibody developed by an inventor.

DEVIL IN THE DETAILS

Over time, however, the technology for antibody analysis has improved. In 2014, two pharmaceutical heavyweights — Amgen in Thousand Oaks, California, and Sanofi in Paris — launched a battle over patents covering a potentially lucrative antibody treatment for high cholesterol.

The case reached a federal appeals court, where judges determined last year that inventors must provide a better description of the actual invention — a more defined set of antibodies — that they wanted to patent.

The USPTO responded with new guidelines for its examiners this year. Since then, patent rejections have piled up. A few weeks ago, patent attorney John Kilyk of Leydig, Voit & Mayer in Chicago, Illinois, learned that an application he was handling had run into trouble. “It was sufficient a few months ago, and now it’s not,” he says.

The court ruling is retroactive, so the move also jeopardizes existing antibody patents. “There’s no doubt that the biotech companies that have been patenting antibodies are

going to be harmed,” says Storz. “There are a number of antibody patents that are now invalid, or would be if somebody tried to enforce them.”

Universities in particular might struggle to put together the information now needed to win an antibody patent, says Rodney Sparks, an attorney with the University of Virginia’s

“There’s no doubt that the biotech companies that have been patenting antibodies are going to be harmed.”

technology-transfer office in Charlottesville. Examiners are asking for more detail about the range of antibodies that can bind to a target and, specifically, where on the target those antibodies will attach. “In universities, our guys want to publish,” Sparks says. “We don’t have the ability, typically, early on to make lots and lots and lots of antibodies and screen for all of those characteristics.” As a result, he says, universities will need to file narrower patents covering only a few of the possible antibodies, and might struggle to find companies willing to license them.

And applicants are facing pushback from patent examiners who are extending the tightened rules on an invention’s written description to other kinds of patent

applications, says Rigby. A broad patent for a method to treat disease by targeting a specific protein, she says, might now also be in question.

“It’s not clear whether examiners have misunderstood and are overreaching, or whether this is a more general trend that the patent office is behind,” Rigby says.

Yet the shift has been an unexpected boon to some companies. Benjamin Doranz, president of Integral Molecular, a company in Philadelphia, Pennsylvania, that produces and analyses antibodies, says that clients used to request analyses mainly to learn more about how their antibodies functioned. But increasingly, he says, the company’s data are being used to bolster patent applications. Some of its clients are now patent-law firms.

Patenting antibodies has become much more treacherous, says Doranz. “But they’re still of great value,” he says, “so everyone is trying to figure out the new patent landscape, and how do we navigate it.” ■

CORRECTION

The World View ‘Transparency rule is a Trojan Horse’ (*Nature* **557**, 469; 2018) misstated the number of signatories to the joint statement. It omitted to mention Cell Press and PLoS journals.



Thousands of Kenyans are taking part in a trial in which they will receive substantial monthly or yearly payments.

THE ANTI-POVERTY EXPERIMENT

Several projects are testing the idea of doling out a 'universal basic income' that people can use however they want.

BY CARRIE ARNOLD

Along the shores of Lake Victoria in western Kenya, mobile phones in several hundred villages ding in unison on the first of every month. For more than 21,000 adults, the sound means one thing: 2,250 Kenyan shillings appearing in their bank accounts. The cash equals one-quarter to half of the average income for a two-adult household in Bomet County, one of the poorest in Kenya.


The money (roughly US\$22.50) arrives courtesy of the US-based charity GiveDirectly, which is studying the effects of handing people lumps of cash with no strings attached — an idea known as a universal basic income (UBI). The mobile phones in these villages will ding every month for the next 12 years, making this UBI trial the longest and largest ever conducted.

"It's a poverty-alleviation tool. Participants can invest in riskier things because they have their basic needs taken care of," says Tavneet

Suri, an economist at the Abdul Latif Jameel Poverty Action Lab at the Massachusetts Institute of Technology in Cambridge and one of the lead investigators on the Kenya trial.

The Kenya experiment is one of a handful of UBI trials in various stages of development around the world. Finland has already begun a trial, as has Ontario in Canada. Stockton, California, is planning to roll out its own experiment later this year. Although the concept isn't new — it was first proposed by Enlightenment philosophers — it remained a fringe idea until

JONAS BENDIKSEN/MAGNUM



the past few years, and governments are now starting to take it more seriously. Interest in the idea grew in the aftermath of the 2008 economic crisis and because of endorsements from Silicon Valley tech gurus such as Elon Musk.

Proponents of guaranteed income schemes argue that poor people will benefit more from unrestricted funds than from current welfare systems, which tend to have stringent requirements that often leave recipients trapped in poverty. “Universal basic income is about giving people cash without question, and trusting that they know how to use it in the most-effective way they can,” says Luke Martinelli, an economist at the University of Bath, UK.

For economists and public-policy scholars, the current interest in UBI provides an opportunity to conduct rigorous trials to determine whether it will produce measurable benefits. But translating a grand economic theory into workable policy is far from easy. Almost all trials have involved a small number of people or lasted just a few years, which limits their power. And there is no clear definition of success; researchers try to balance measuring potential gains in one area, such as health care, with potential offsets in another, including education and labour-force participation.

But for the growing chorus of voices calling for data-driven policy, trials such as the one in Kenya are the only way to see whether UBI actually works. “This is one of the first rigorous randomized control trials of UBI,” says Suri. “This is our chance to understand UBI and its impacts.”

NO-STRINGS CASH

The modern welfare state emerged out of the ashes of the Great Depression and the Second World War. As governments across the Americas, Europe and Commonwealth tried to rebuild their economies, they began taking an active role in providing for the well-being of their poorer citizens through grants, services and money earmarked for purposes such as housing and food. Although such welfare systems have improved standards of living, most require an immense bureaucracy to administer benefits and to ensure that recipients meet strict qualification standards.

Welfare critics have long argued that the administrative costs are huge and provide limited positive results; in some cases, they discourage people from finding jobs. In response, leaders across the political spectrum have latched onto the idea of UBI — which has been promoted over the centuries by luminaries such as Thomas More (in his 1516 novel *Utopia*), philosopher Thomas Paine, the liberal US President Franklin Delano Roosevelt and economist Milton Friedman, a favourite of conservatives including US President Ronald Reagan and UK Prime Minister Margaret Thatcher. Progressive politicians and thinkers

have seen the idea as a way to end poverty; conservatives have viewed it a streamlined welfare system that is easier and cheaper to run.

In the 1960s and 1970s, a handful of sites across the United States tested a scheme related to UBI called a negative income tax. In this kind of programme, individuals making below a certain amount receive supplemental money from the government. But after early results from one of the trial sites revealed an increase in divorce rates, politicians nixed the idea as being toxic to the American family.

“Universal basic income is about giving people cash without question.”

Another early test happened across the border in the small prairie town of Dauphin, Canada. In a trial called Mincome, supported by the federal and provincial governments, the town's poorest residents received monthly cheques from 1974 to 1978 with no constraints on how the money should be spent. Researchers tracked changes in the proportion of people working full- and part-time, as well as in nutrition, education and basic health outcomes. But before the trial could be analysed, waning funds and political change scrapped the idea, and all the data were packed in more than 1,800 boxes and stored in a warehouse. They sat there until economist Evelyn Forget at the University of Manitoba in Winnipeg brushed off layers of dust and opened the boxes.

The documents Forget uncovered revealed that teenage children in MINCOME families completed an extra year of schooling compared with teens in similar small Manitoba towns. Hospitalizations decreased by 8.5%, with the largest drops in admissions for accidents and injuries and mental-health diagnoses. Importantly for economists, who worried that the programme might encourage people to quit their jobs, Forget found that employment rates stayed the same throughout the trial (E. L. Forget *Can. Public Policy* 37, 283–305; 2011).

Now, supporters of UBI in several countries are trying to build on the results from those earlier studies and develop trials to decide whether governments should give UBI a chance.

READY MONEY

The Kenyan experiment grew out of smaller trials that the charity GiveDirectly had done in sub-Saharan Africa. Starting in 2009, the group tried to alleviate poverty by providing relatively modest direct cash transfers. These shorter and smaller injections of cash created ripple effects through the communities

involved. In a trial in Zimbabwe, one year of cash transfers improved childhood vaccination rates and school attendance (L. Robertson *et al. Lancet* 381, 1283–1292; 2013). Because the transfers were only short-term and too small to cover living expenses, they weren't full-blown UBI trials. But that early work gave the group a leg up on planning a UBI trial, says Michael Faye, co-founder and president of GiveDirectly.

Experts say that full-blown experiments are particularly difficult to design because they require a great deal of advance planning.

“With a lot of these projects, the devil is in the details, and the design of research depends on a fine-grained knowledge of its impact,” says Rob Reich, a political scientist at California's Stanford University, who is not part of these trials.

Guaranteed-income experiments are different from many clinical trials because researchers are looking for improvements in a wide variety of areas and doing so across communities rather than in individuals. Suri describes a cycle of improvements that UBI might create: well-fed pregnant mothers should have healthier children than would undernourished women. Longer education would create better job opportunities and delay marriage and childbirth, resulting in healthier mothers and babies. Suri says that her team plans to track everything from entrepreneurship to health, education and nutritional status, with the help of a platoon of locals who will go door-to-door, and do a series of short phone check-ins and some in-depth interviews with village elders to get a big-picture view of the intervention's effects.

Because the trial will be so long and expensive, Suri helped to design four different arms to answer as many questions as possible. The largest arm provides 2,250 Kenyan shillings every month for 2 years to each adult in 80 villages. A second arm provides the same amount of money each month for 12 years. A third arm provides a total equivalent to US\$505 — 2 years' of basic income — in 2 payments, separated by 2 months. The fourth arm serves as a control group that gets nothing.

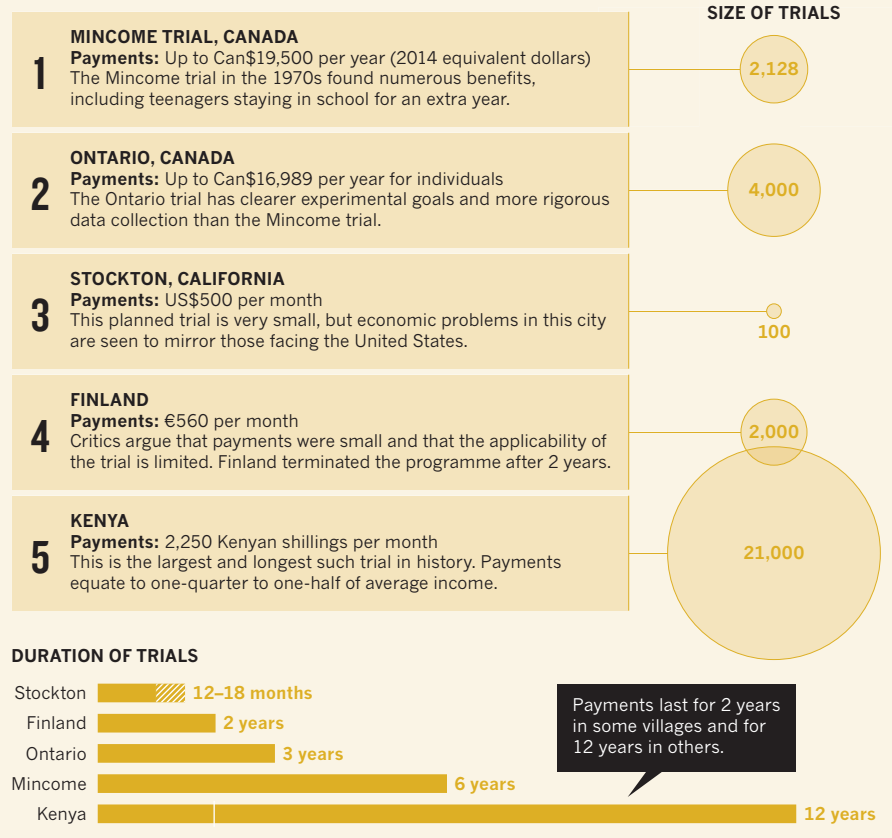
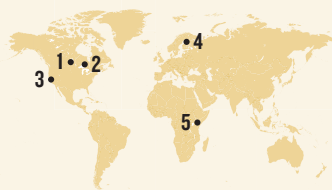
“We can run a horse race between different types of UBI,” Suri says.

Participants in a pilot project that began in 2016 are enthusiastic about their prospects. “This has made me believe that I can commit and be able to pay school fees for my children and I am also confident of saving money to improve my business,” says Jael.

A UBI experiment in Finland has been struggling. The project grew out of concerns that the country's complicated unemployment benefit system keeps some people from returning to full-time work because that would curtail their support. In March 2016, the government welfare agency Kela teamed up with a non-profit research organization Tänk to announce a UBI trial that would provide €560 (US\$658) per month to a group of 2,000 adults currently receiving ►

MONEY FOR NOTHING

Canada conducted an early test of universal basic income in the 1970s. Several governments and charities are now experimenting with other schemes.



► unemployment benefits around the country. The extra income would not be taxed at the same rate as normal unemployment benefits, nor would participants be required to actively search for work during the two-year trial. They also wouldn't lose the UBI income if they found work.

Although global media initially praised the programme, popular opinion later soured on the scheme, which cost €20 million. The monthly payments were nowhere near enough to cover an adult's basic living expenses, and the trial addressed only adults who were unemployed at the time. Plus, there was no adequate control group. In late April 2018, the Finnish parliament refused requests from Kela for another year of funding and instead expressed preferences towards developing other welfare schemes. To UBI advocates, the problems with the Finnish programme have proved a serious barrier to getting other trials up and running.

"People's expectations were far higher than the trial could deliver," says Karl Widerquist, a political economist at Georgetown University in Qatar and co-chair of the Basic Income Earth Network, which promotes UBI.

Other researchers have started a range of UBI trials and precursor projects (See 'Money for nothing'). The city of Stockton has begun an experiment that is underwritten by philanthropic organizations. The trial will be smaller than the Kenya trial — including only 100 people for 12–18 months — owing to funding limitations, says Taylor Jo Isenberg, managing director at the Economic Security Project, who is helping to fund the experiment. But the effort will provide good preliminary data for later trials, she says, because Stockton is a microcosm of the United States in terms of diversity, poverty levels and job loss from automation and outsourcing.

"Updating previous research on unconditional cash transfers is really important but really expensive. We hope to open the door for other stakeholders to step in later," Isenberg says.

The Canadian province of Ontario is also hopping on the UBI bandwagon. Its trial began late last year, enrolling more than 4,000 individuals across the province. As spectators watch for signs of success and failure in these trials, the researchers involved need to define what, exactly, 'success' or 'failure' means. Given

that some of the impacts of UBI won't be felt for 5–10 years, onlookers might be waiting for quite some time.

The announcement in April that the Finnish UBI trial wouldn't be funded beyond this year provided a sobering reminder that politics — more than data — will determine the fate of such programmes. The government pulled the plug before Markus Kanerva, managing director at Tänk, and his colleagues had examined the data to see how well the trial worked, something Kanerva says his team is waiting to do until late 2018 so as not to bias the results.

The outcome of all these trials is far from clear, not least because the Kenya project — the most ambitious — only just began.

PROBLEMS OF SCALE

Over time, the trials could generate data on the costs and benefits of UBI schemes, such as whether the initiatives reduce health-care expenditures. But Martinelli thinks that the data will show that it will cost too much to make a programme effective. "An affordable UBI is inadequate, and an adequate UBI is unaffordable," he says.

But even a clear win in these trials won't necessarily indicate that UBI would work in practice, says economist Damon Jones at the University of Chicago in Illinois. Because they are relatively small and most of the funding comes from private sources, the trials won't provide a sense of whether governments could afford a big public programme or whether citizens would be willing to fork out extra taxes to fund them. "Medicine can be scaled up, but this isn't as easy," says Jones. A new cancer drug might extend lifespan by 3 months, which stays true whether 10 people take the drug or 10,000. In a UBI trial, 10 people receiving cash will have a very different impact on a community compared with 10,000.

Jones cautions that this doesn't mean the UBI trials shouldn't be done or that they will produce meaningless data, just that even the best-designed studies have inherent limitations.

Regardless of the outcomes, the trials will have an ongoing impact because they can identify potential flaws in the process, help researchers refine the questions they ask and give policymakers some of the answers they crave. If the trials succeed, "it wouldn't just be an outlier in social policy, it would be a minor miracle," Reich says.

For the participants of the Kenya trial, that minor miracle has already arrived. The knowledge that GiveDirectly will deposit funds into their accounts every month for more than a decade has already begun to shift how some of them think about money. Each text alert means a chance to invest in their own lives or their businesses with the security that they can still put food on the table. And that, they say, is priceless. ■

Carrie Arnold is a freelance journalist in Richmond, Virginia.

COMMENT

NEUROSCIENCE The case for inflammation as a cause of depression **p.633**



HISTORY Why Aldous Huxley's post-nuclear dystopia feels newly relevant **p.634**

FUTURE OF WORK Microbes might end up taking the jobs that robots leave us **p.637**

OBITUARY Peter Grünberg, Nobel prizewinner for data-storage nanotechnology **p.638**

ERIC RUDD/IU COMMUNICATIONS



Michelle Lollie is an American Physical Society Bridge Fellow at Indiana University in Bloomington.

Making physics more inclusive

Theodore Hodapp and **Erika Brown** explain how the American Physical Society is helping to recruit and retain PhD students from under-represented minorities.

African Americans, Hispanic Americans and Native Americans make up about one-third of university-age citizens in the United States. Yet less than 11% of bachelor's degrees in physics are awarded to people from these groups. At the doctoral level it is even worse, with only about 7% of physics PhDs granted to US citizens from racial and ethnic minority groups — just 60–70 students each year. This is one of the lowest rates in the sciences. Chemistry, by comparison, awards 17% of bachelor's and 11% of doctoral degrees to these groups (see 'Doctoral dearth'). The proportion in physics has barely risen over the past 15 years, while the percentage of US university-age students from minorities has grown by 18%.

This is morally questionable and disastrous from a practical point of view. The discipline of physics, and society as a whole, are missing out on talent. Students are often judged on the prestige of their undergraduate institution or the preparation they received at school, rather than on what really matters: their aptitude, drive and ingenuity.

Physicists cannot fix all of society's ills, but the community can and must provide more equitable pathways into research. This does not mean lowering the bar, but showing students where it is and helping them to find their way over it.

For the past five years, the American Physical Society (APS) has been taking the first steps by working with physics departments across the United States to balance the doctoral and bachelor's graduation rates for under-represented students. Given that the numbers of students are small, interventions at a limited number of universities can drastically change the landscape. To effect this change, the APS has directed resources to overcoming admissions barriers and ensuring that graduate programmes where students are admitted have adequate support to help them remain on track. These support structures benefit all students.

The APS Bridge Program¹ (funded in part by the US National Science Foundation) asks physics faculty members to consider and recruit graduate students from under-represented minorities whom they think would do well in a doctoral programme but who, for whatever reasons, have not ▶

► been accepted. Such recommendations are permitted, although it is illegal in the United States to specify race or ethnicity in university admissions procedures as the sole criterion for a decision.

After the standard mid-April cut-off for informing students of their acceptance into US graduate programmes, the APS collects applications from Bridge Program candidates and circulates them to institutions. The institutions take another look and select the students who are best for them. The departments are required to mentor and monitor the progress of Bridge Program students. More than 35 US institutions are now working with the APS.

There are currently around 150 students in the Bridge Program. In 2017, by accepting 46 students in one year, departments more than compensated for the difference between the doctoral and bachelor's graduation rates (see 'Bridging the gap'). When the APS began the programme in 2012, it gave grants to universities to support most Bridge students. Now, most students are funded by the physics community; in 2017–18, the APS supported only six.

We found no single root cause for why under-represented students were not accepted into graduate programmes in physics. The problems were mostly systemic and circumstantial, not the fault of the students. Some students told us that they were unable financially to apply to more than a few programmes, or that they were discouraged by perceived and real biases in application procedures. Other factors included inadequate mentoring and preparation for research careers at the student's undergraduate institution. These hindrances are relatively easy to overcome.

Here we discuss what we've tried, what we have found to work and what still needs to be explored.

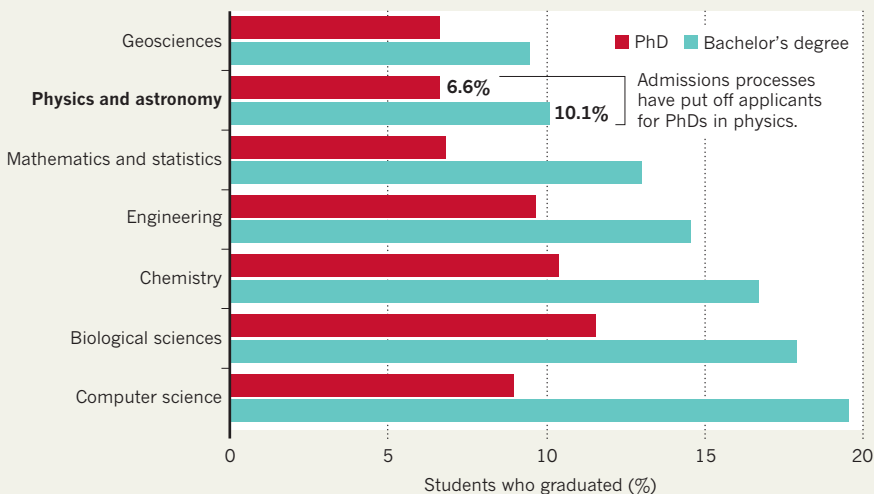
GRADUATE ADMISSIONS

The first hurdle is the graduate admissions process. It is a well-guarded door lying between a student and a research career. Committees must contend with hundreds of applications and an incomplete picture of each student. Candidates with high scores in undergraduate mathematics and physics courses or entrance exams pass through the door easily, including some students from minority groups. Applicants who have mixed academic records can benefit from further consideration by admissions committees.

Behind each CV is a story. What if you went to a substandard middle or high school, where your peers barely made it through algebra and the teacher taught far below your potential? What if you had to find a full-time job to finance your university education, leaving little time to study, much less excel? Some students in

DOCTORAL DEARTH

In all disciplines across the sciences, the proportion of US citizens from under-represented minorities graduating with bachelor's degrees is low; the proportion completing PhDs is even lower.



SOURCES: APS/PPDS COMPLETION SURVEY

our programme experienced these situations. Remedies were as simple as extra coursework to compensate for inadequate preparation, a graduate stipend to provide financial stability, or a committee that was able to see past one poor mark to recognize potential.

In our experience, the biggest barrier to students getting into a physics doctoral programme is the Graduate Record Examination (GRE), a standardized test required for admission into most graduate schools in the United States. More than one-third of US graduate physics programmes will consider only candidates whose scores in the physics GRE test (P-GRE) exceed a cut-off². This ignores the larger picture of a student's development and also goes against the advice of the Educational Testing Service (ETS), which produces the GRE. The ETS recommends that GRE scores should never be the sole basis for an admissions decision and should be weighed against other factors³.

P-GRE scores conflate many things. Students need to prepare carefully because the scope and approach of the test are different from how most undergraduates are taught and evaluated. In addition, many undergraduate institutions offer no tools or guidance to help students to prepare. It costs US\$150 to take the test. Despite the best efforts of the ETS, the GRE tests suffer from biases resulting from students' societal experiences and expectations. Women and people from minority racial and ethnic groups score lower than do white or Asian men, on average⁴. Candidates are

“Issues unrelated to academic ability can affect or destroy a graduate student's potential to stay the distance.”

influenced by 'stereotype threat': members of groups for which stereotypical expectations are low perform worse in high-stakes exams when they are reminded that they are part of that group (see, for example, ref. 5). These factors, and a student's aptitude for taking this type of test — or even how well they were feeling on the day — matter.

Scientists should care most about potential, not preparation. Even if admissions committees downplay the value of the GRE, students do not. Those with low scores are discouraged from applying to institutions that publish high average scores.

The question then remains, how should admissions committees pick graduate students?

This is both a philosophical and a practical concern: what are committees' goals in selecting a student, and how should they sort through a big pile of applications in a small amount of time?

Philosophically, should committees try to identify the student who is already at the top of the applications pile, itself defined in part by systemic biases? Or should they try to spot someone who can develop to become an excellent researcher? The latter mindset⁶ accommodates individuals who might have grown up in places with few educational and mentoring resources available, but who have a passion and aptitude for physics. Members of the physics community should provide an opportunity for such individuals, irrespective of their social background.

Practically, the APS works with departments that are trying a variety of ways to select students. It's too soon to tell how these strategies can be generalized. Each department has different needs and must find a technique that works for it. Some review all applications from target groups to find



APs Bridge Fellow Joseph (JB) Holmes is studying biological physics at Indiana University, Bloomington.

compelling stories that indicate promise. Others shortlist potentially good candidates on paper and conduct short, 15–30-minute video interviews with each. These explore traits that are correlated with success, such as problem solving, tenacity and the ability to assess your own weaknesses (see, for example, ref. 7).

STUDENT SUPPORT

The next step is to help PhD students to finish their doctorates. Mentoring and peer support are crucial.

All graduate students face challenges. In a 2008 study, only 59% of US doctoral students in physics completed their PhDs⁸. As well as missing out on talent, it is expensive to lose graduate students. Each requires upwards of \$300,000 of direct support during their studies, as well as resources, facilities and faculty members' time. Students are committing years of their life towards the long-term goal of engaging with the physics community.

The Bridge Program, by contrast, has an average retention rate of around 85%. How have institutions done it?

Interviews with Bridge students and their mentors have revealed that numerous practical issues unrelated to academic ability can affect or destroy a graduate student's potential to stay the distance. Examples include: living too far from campus to join in study or research sessions; being inexperienced in managing money; family commitments

and dynamics; feelings of isolation; or poor advice on how to navigate the university system. Poverty exacerbates all these problems.

Several mentors are preferable, including a research adviser, an academic adviser and someone whom the student feels has no power over them, such as a staff member. Bridge students check in with their

mentors at least once every couple of weeks during the first year so mentors can make sure they are adjusting well. Meetings can taper off as students find their groove. But it is important that mentors intervene early when problems arise, such as illness, personal issues or courses that are pitched at an inappropriate level.

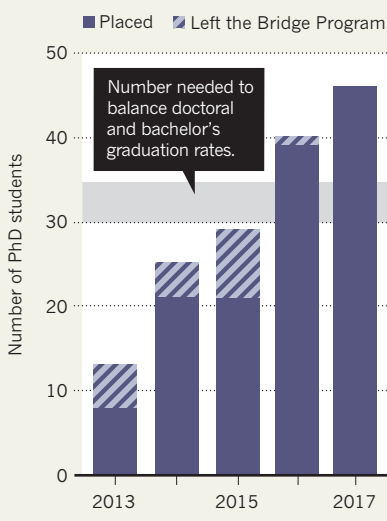
We have found the first six weeks to be crucial. Changes to a student's academic plan after this come too late — students facing obstacles already feel that pursuing graduate education was a bad idea; isolation has set in. They might already be well down a downward spiral that leads to dropping out.

Peer support is crucial, too. Institutions involved with the Bridge Program either had or have developed a physics graduate student association. These work on behalf of all students, but their activities can be pivotal for students from diverse backgrounds who are feeling isolated. The student associations assign more-senior students as mentors to new participants in the Bridge Program, hold social functions to welcome all students, and provide a space for them to share experiences and knowledge. Some hold student-only seminars — at which no faculty members are allowed — on careers, courses and campus life, providing a place to vent and learn. Representatives of these organizations can be a 'student voice' in conversations with the faculty.

Problems can and do often occur late

BRIDGING THE GAP

In its fourth year, the American Physical Society's Bridge Program admitted enough students to erase the difference between graduation rates for bachelor's and PhD degrees in physics.





JOSHUA SWEET

Physics PhD student Keanna Jardine is investigating the dust dynamics of small asteroids at the University of Central Florida, Orlando.

in a student's studies as they navigate the research and dissertation phases. In many universities, a committee meets annually to review the progress of each graduate student. Ideally, the chair of such a committee should be someone other than the student's research adviser, in case that relationship sours. Faculty members might need to devote extra mentoring time at this stage to ensure the student finishes their work and thesis.

The APS tracks all students in the Bridge Program. Along with academic transcripts, we ask mentors to evaluate each student's progress towards a PhD. The proof is in the retention rate: currently, 85% of our students are on track — significantly more than the national average. Students report that the programme gave them the chance they needed to pursue graduate studies.

Our first students are likely to receive their PhDs in 2019. They will then start looking for postdoctoral jobs. The APS has begun to collaborate with national laboratories in the United States — collectively the largest employer of physicists outside academia — to help match up Bridge Program graduates with job posts. We are also developing a mentoring curriculum for the researchers who sponsor these graduates,

"The world cannot afford to waste talent."

to help make them more aware of diversity issues.

NEXT STEPS

To make the physics community more representative, we recommend three actions.

First, graduate departments should aim to reflect the racial, ethnic and gender mix of the undergraduate population pool, at a minimum. They should use admissions techniques that look beyond conventional measures, to identify students who can be successful leaders in the future⁹. Admissions committees must educate themselves on what the P-GRE is actually measuring, rather than what they think it is measuring.

Second, graduate departments should foster more-supportive cultures for all students. Departments should offer undergraduate coursework where needed, mentor students throughout their studies — especially in the first few semesters — and formalize mentoring by peers.

Third, we encourage other national organizations, such as the American Chemical Society, the American Geophysical Union and their equivalents in other countries to take a similar intermediary role. We have begun discussions with some of these and received enthusiastic responses. Moreover, similar interventions could reduce gender disparities in disciplines in which the percentage of women changes appreciably

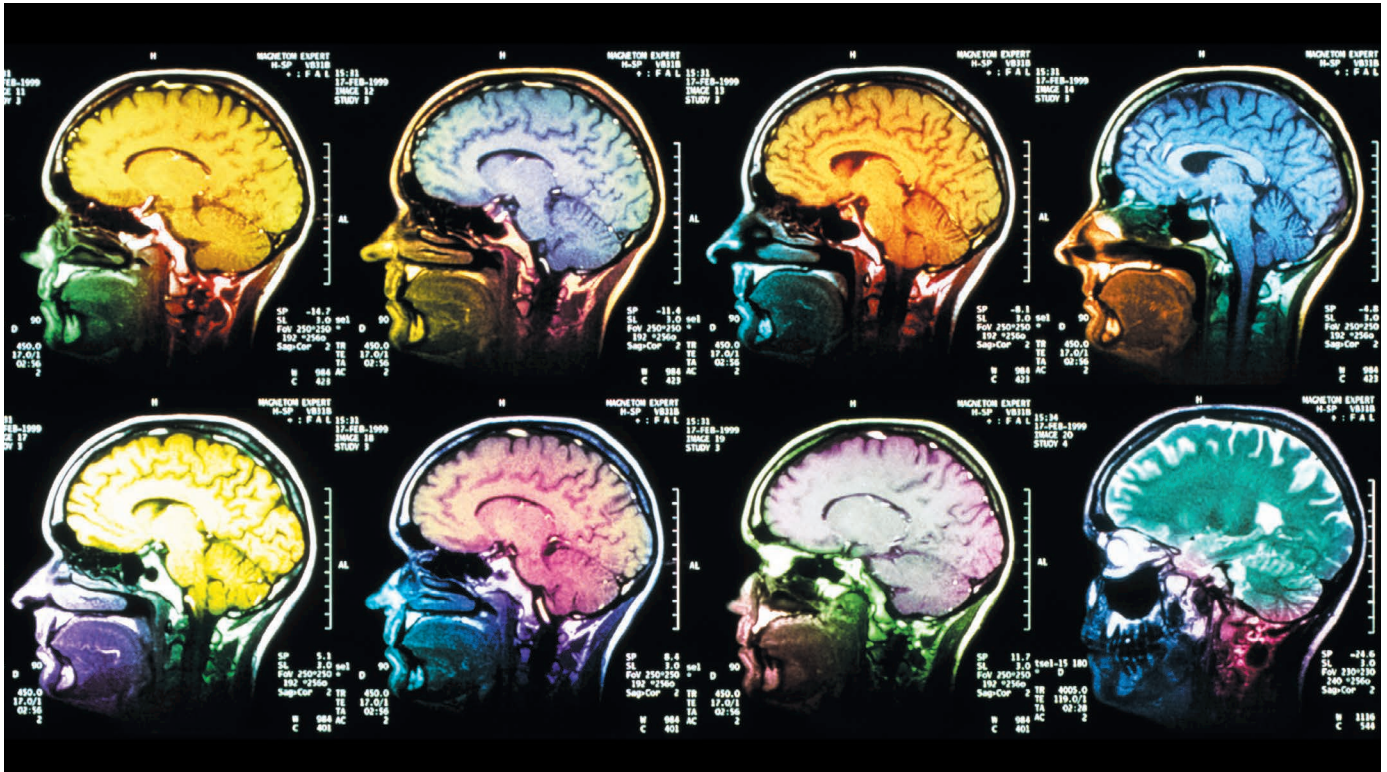
between undergraduate and graduate stages (in physics it does not).

We must embrace diversity within the physics community. The world cannot afford to waste talent. ■

Theodore Hodapp is director of project development and senior adviser to the Education and Diversity Department at the American Physical Society, College Park, Maryland, USA. **Erika Brown** is Bridge Program manager at the American Physical Society.

e-mail: hodapp@aps.org

1. Hodapp, T. & Woodle, K. *Phys. Today* **70**, 2, 50 (2017).
2. Potvin, G., Chari, D. & Hodapp, T. *Phys. Rev. Phys. Educ. Res.* **13**, 020142 (2017).
3. Educational Testing Scores. *GRE Guide to the Use of Scores* (ETS, 2017).
4. Miller, C., Zwickl, B., Posselt, J., Silvestrini, R. & Hodapp, T. *Sci. Adv.* (in the press).
5. Steele, C. M. *Whistling Vivaldi: And Other Clues To How Stereotypes Affect Us* (Norton, 2010).
6. Scherr, R. E., Plisch, M., Gray, K. E., Potvin, G. & Hodapp, T. *Phys. Rev. Phys. Educ. Res.* **13**, 020133 (2017).
7. Duckworth, A. L., Peterson, C., Matthews, M. D. & Kelly, D. R. *J. Pers. Soc. Psychol.* **92**, 1087–1101 (2007).
8. Sowell, R., Zhang, B. & Redd, K. *Ph.D. Completion and Attrition: Analysis of Baseline Demographic Data from the Ph.D. Completion Project* (Council of Graduate Schools, 2008).
9. Posselt, J. R. *Inside Graduate Admissions: Merit, Diversity, and Faculty Gatekeeping* (Harvard Univ. Press, 2016).



Magnetic resonance imaging scans of the human brain.

PSYCHIATRY

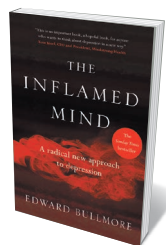
Depression revisited

Alison Abbott considers a persuasive case that inflammation is linked to the disorder.

Depression affects one in four people at some time in their lives. It is often difficult to treat, in part because its causes are still debated. Psychiatrist Edward Bullmore is an ardent proponent of a radical theory now gaining traction: that inflammation in the brain may underlie some instances. His succinct, broad-brush study, *The Inflamed Mind*, looks at the mounting evidence.

The book outlines a persuasive case for the link between brain inflammation and depression. Bullmore pleads with the medical profession to open its collective mind, and the pharmaceutical industry to open its research budget, to the idea. He provides a current perspective on how the science of psychiatry is slowly emerging from a decades-long torpor. He sees the start of a shift in the Cartesian view that disorders of the body 'belong' to physicians, whereas those of the more 'immaterial' mind 'belong' to psychiatrists. Accepting that some cases of depression result from infections and other inflammation-causing disorders of the body could lead to much-needed new treatments, he argues.

In 1989, during his clinical training at St Bartholomew's Hospital in London, Bullmore encountered a patient whom he



The Inflamed Mind: A Radical New Approach to Depression
EDWARD BULLMORE
Short (2018)

calls Mrs P, who had severe rheumatoid arthritis. She left an indelible impression. He examined her physically and probed her general state of mind. He reported to his senior physician, with a certain pride in his diagnostic skill, that Mrs P was both arthritic and depressed. Replied the experienced rheumatologist dismissively, given her painful, incurable physical condition, "You would be, wouldn't you?"

Mrs P is a recurring motif, as is the rhetorical question. Bullmore draws on more than two millennia of medical history — from ancient Greek physician Hippocrates to the work of neuroanatomist and 1906 Nobel laureate Santiago Ramón y Cajal — to illustrate his points. At times they seem like intellectual meanderings, but these passages also show how medical science often progresses by means of bold theories

that break away from received wisdom.

After his training, Bullmore specialized in psychiatry, and quickly experienced its limitations. He describes his growing awareness of how poorly science has served the field, using the development of selective serotonin reuptake inhibitors (SSRIs) as a prime example.

That long and winding road began with the antibiotic iproniazid. It was discovered through scientific logic: by screening chemicals for their ability to kill *Mycobacterium tuberculosis* in the test tube and in mice. Iproniazid transformed the treatment of tuberculosis in the 1950s. Patients clawed back from the jaws of death exhibited euphoria — well, you would, wouldn't you? — and the drug was soon launched as an antidepressant. Soon the theory emerged (based more on supposition than evidence, says Bullmore) that its psychiatric effects were the result of boosting the neurotransmitters adrenaline and noradrenaline. Drug developers began to focus on neurotransmission more broadly.

Prozac (fluoxetine), which boosts serotonin transmission, was launched in the mid-1980s, and many pharmaceutical

► companies quickly followed with their own SSRIs. It seemed to be the revolution psychiatrists had been waiting for. But it soon emerged that only a modest subset of patients benefited (estimates based on trials vary widely). That is unsurprising in retrospect, with the new appreciation that depression can have many causes. Bullmore holds that the emergence of SSRIs bypassed scientific logic. The serotonin theory, he writes, is as “unsatisfactory as the Freudian theory of unquantifiable libido or the Hippocratic theory of non-existent black bile”. He notes that, after SSRIs failed to live up to the hype, time once again stood still for psychiatry.

Bullmore recalls a teleconference in 2010, when he was working part-time with British pharmaceutical giant Glaxo-Smith Kline. During the call, the company announced it was pulling out of psychiatry research because no new ideas were emerging. In the following years, almost all of ‘big pharma’ abandoned mental health.

Then a window seemed to open — one that shed a different light on the plight of Mrs P. Some of the textbook certainty that Bullmore had learnt by rote at medical school started to look distinctly uncertain.

In particular, the blood-brain barrier turned out to be less impenetrable than assumed. A range of research showed that proteins in the body could reach the brain. These included inflammatory proteins called cytokines that were churned out in times of infection by immune cells called macrophages. Bullmore pulls together evidence that this echo of inflammation in the brain can be linked to depression. That, he argues, should inspire pharmaceutical companies to return to psychiatry.

It seems unfair that someone struck down by infection should have depression too. Is there a feasible evolutionary explanation? Bullmore hazards that depression would discourage ill individuals from socializing and spreading an infection that might otherwise wipe out a tribe.

Other brain disorders might turn out to be prompted or promoted by inflammation. An exciting link with neurodegenerative diseases, including Alzheimer’s, is also being studied (see *Nature* 556, 426–428; 2018). But we need to learn from the rollercoaster history of brain research, and keep expectations in check. Beneath his bombastic enthusiasm, Bullmore acknowledges this, too. ■

Alison Abbott is *Nature’s* senior European correspondent.

“After SSRIs failed to live up to the hype, time stood still for psychiatry.”

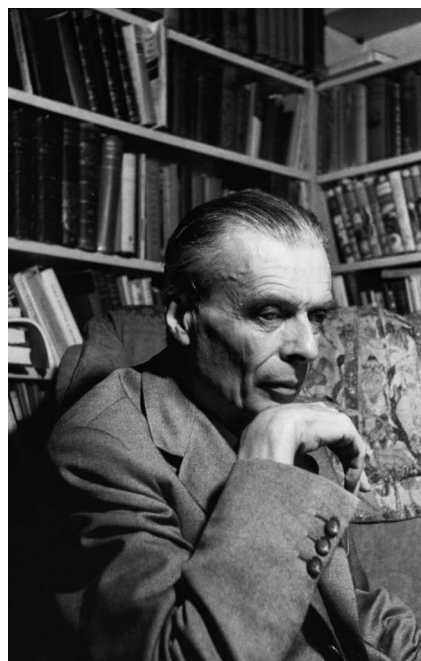
IN RETROSPECT

Ape and Essence

Richard Rhodes finds resonance with today’s uneasy nuclear age in Aldous Huxley’s satirical dystopia.

The atomic bombings of Hiroshima and Nagasaki in August 1945 shocked the world. More than 70 years on, these events have not been repeated — evidence that it was the United States’ temporary nuclear monopoly that made them possible. Yet few observers in the immediate postwar years foresaw the development of an uneasy nuclear truce, enforced by the certainty of mutual destruction. Fears of an arms race culminating in nuclear war were widespread. Such fears have now re-emerged, with North Korea’s burgeoning arsenal and the United States’ abrogation of its agreement with Iran.

J. Robert Oppenheimer, who directed the development of the first bombs, was among those who feared nuclear conflict, and worked for international control after the Second World War. Physicist Richard Feynman (whom I interviewed for my 1987 book *The Making of the Atomic Bomb*) recalled sitting in a bar in New York in 1946, watching the crowd passing outside and thinking: “You poor fools, you have no idea that in a few more years you’ll all be dead.” Aldous Huxley seems to have leapt to the same conclusion in his hybrid novel and film scenario *Ape and Essence*, published 70 years ago.



Aldous Huxley in the 1950s.

Ape and Essence

ALDOUS HUXLEY
Harper & Brothers
(1948)

The prolific novelist and essayist had been formulating his thoughts on the bomb since the end of the

war. In 1947, Huxley published the extended essay ‘Science, Liberty, and Peace’, a prelude to the novel. There, he wrote that the power-hungry and nationalistic “boy-gangster” in us all would easily prevail over the reasonable adult, exulting: “Press a few buttons and bang! the war to end war will be over, and I shall be the boss of the whole planet.” Huxley knew better. If more than one nation had such weaponry, he believed, the outcome of “the war to end war” would be world-scale destruction. And because that would be a kind of singularity, it seemed to him that almost anything might follow.

Ape and Essence is Huxley’s imagining of a post-nuclear world. The title is from William Shakespeare’s *Measure for Measure*: Isabella speaks of the proud man’s “glassy essence, like an angry ape”, which “plays such fantastic tricks before high heaven/As make the angels weep.” The angels have flown in Huxley’s novel, set in what remains of Los Angeles, California, in 2108 — a century after a third world war, which would have taken place around now, in Huxley’s fictional timeline.

In one of the book’s set pieces, intelligent baboons fight this twenty-first-century war, with scientific luminaries (Michael Faraday and two opposing Albert Einsteins) as leashed mascots. So much for scientists, Huxley insinuates — “good, well-meaning men, for the most part. But ... they ceased to be human beings and became specialists.” Of the two opposing cultures described by scientist and novelist C. P. Snow in 1959, Huxley was clearly on the side of the humanities, as if locked in debate with his distinguished scientific kin — his biologist brother Julian, physiologist half-brother Andrew and zoologist grandfather Thomas Henry, known as Darwin’s bulldog.

Having introduced the baboons, Huxley kills them off: it’s a second false start to the novel’s stuttering story, a metafictional concoction as multilayered as an onion. The first storyline sees two screenwriters tracking down a legendary colleague, only to find him dead. The deceased’s abandoned screenplay (I suspect one of Huxley’s own unsold efforts, repackaged) is the book’s centrepiece. It is

AGE FOTOSTOCK/ALAMY

here that the baboons rise and fall. The script then moves on to an improbable love story set in the world of an irradiated rump of humans who survived the war but have forgotten how to make things. They live by scavenging leftovers from the pre-war days, burning books for heat and assigning crews to rob old graves of suits and jewellery. Eunuch priests of the devil-figure Belial squat at the top of the caste system in this stunted world, dominating a society of near-slaves.

Ape and Essence parallels Huxley's 1932 *Brave New World* (see P. Ball *Nature* **503**, 338–339; 2013), yet offers an even darker vision. A young botanist, Alfred Poole, has arrived by ship from New Zealand, which survived the atomic war and is now exploring what's left of the world. So, as with *Brave New World's* Savage, a *Candide*-like hero appears from outside society and finds himself appalled. And what appals both heroes is the indiscriminate sexuality that the society's leaders encourage to replace family and human love.

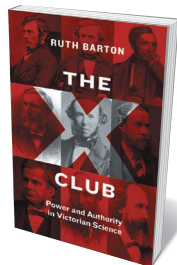
The twist this time, as Huxley wrote to his fellow screenwriter Anita Loos, is that “the chief effect of the gamma radiations [has] been to produce a race of men and women who don't make love all the year round, but have a brief mating season”. This manifests as mass gropings; any progeny deemed too monstrous, the result of radiation-damaged genes, are then slaughtered on Belial Eve. (Huxley probably knew that Hermann Muller had received a Nobel prize in 1946 for the discovery that X-rays can cause mutations.) The ceremony, called the Purification of the Race, mimics the blood sacrifices of the Aztecs. It also alludes to eugenics, the British–American pseudoscience embraced by Adolf Hitler.

What all this sexualized barbarity has to do with nuclear war isn't clear. Born in 1894, Huxley brought a scolding Victorian sensibility to the loosened morals of the war-torn twentieth century, excoriating its hedonism in satires and science fiction. Sun-drenched, beauty-obsessed southern California, where he lived and worked from 1937 until his death in 1963, proved the ideal locale for his dystopias: a seeming paradise that was also the end of the frontier.

Appropriately enough, *Ape and Essence* culminates in a Hollywood happy ending, at least for Poole and Loola, the young woman he falls in love with. The lovers escape to northern California, where a colony of “hots” — hold-outs with conventional sexuality — are cobbling together a new life. However disdainful Huxley might have been of our core boy-gangsters, in the end, he was too humane for a truly relentless apocalypse: his dystopias had escape hatches. Would that the same could be said of a real nuclear war. ■

Richard Rhodes's latest book, *Energy: A Human History*, will be published in the United States in late May.
e-mail: richardrhodes1@comcast.net

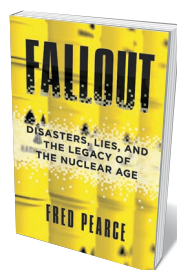
Books in brief



The X Club

Ruth Barton UNIVERSITY OF CHICAGO PRESS (2018)

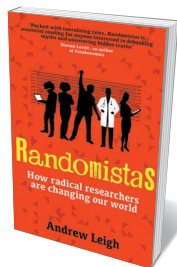
For decades in the late 1800s, nine scientific luminaries (among them biologist Thomas Henry Huxley and botanist Joseph Dalton Hooker) dined together as members of the ‘X Club’. This socio-economically diverse group, formed in part to promote Charles Darwin's achievements, is a telling case study in the dynamics of Victorian class and science. Historian Ruth Barton's magisterial chronicle traces the careers of the “X-men” and their agile promotion of science; Huxley, in particular, emerges vividly as wily, belligerent, and obstructive to women entering science.



Fallout

Fred Pearce BEACON (2018)

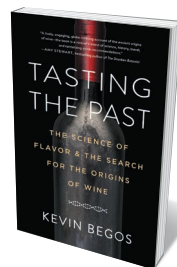
Science writer Fred Pearce casts a cool and measured eye on an explosive legacy: the atomic age. Launched by Winston Churchill's nuclear ambitions (realized by the US Manhattan Project), this era lingers on in plutonium stockpiles, arsenals and ageing power plants. Pearce roams with intent from Sellafield, “Britain's brooding nuclear nightmare”, to radioactive steppes in Kazakhstan, blighted by 619 atomic tests in the 1950s. His nuanced conclusion is that, together, alarmist protestors and a secretive nuclear industry create a different sort of fallout: the spread of disinformation and fear.



Randomistas

Andrew Leigh YALE UNIVERSITY PRESS (2018)

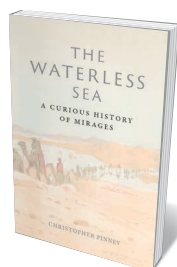
Randomized testing, economist Andrew Leigh reminds us, has vanquished scurvy, improved wildfire response — and proved key to better feedback loops in medicine and crime prevention. The trove of case studies in his insightful study includes the 1960s Perry Preschool Project, which exposed the long-term positive impact of early education among African American children living in poverty. Leigh also explores the work of pioneering ‘randomistas’ such as social-policy expert Judith Gueron, and outlines handy guidelines on aspects of randomized testing, such as sample splitting and ethical oversight.



Tasting the Past

Kevin Begos ALGONQUIN (2018)

If you can tell Sauvignon blanc from Sémillon, you might feel that you ‘know’ wine. Science journalist Kevin Begos blows that idea to smithereens. He travelled from the Caucasus Mountains to Israel and beyond, and rifled through archives, to unearth ancient ‘founder’ grape varieties. En route, he consults archaeobiologist Patrick McGovern and grape geneticist Shivi Drori; reads papers on the DNA of “wild yeasts that live symbiotically with wasps”; and contemplates the oldest grape fossil found. A book that froths with data on half-forgotten vines, from Hamdani to Gros Manseng.



The Waterless Sea: A Curious History of Mirages

Christopher Pinney REAKTION (2018)

The illusory seas observed in sere deserts are not the only form of mirage, notes Christopher Pinney in this alluring tour of the phenomenon in science and culture. Created by light refracting as it moves through atmospheric regions with differing temperatures, mirages can also appear as imposing and mysterious ‘castles in the air’. Pinney ranges from the old Japanese belief that these “phantom paradises” were exhaled by clam monsters, to an 1898 *Nature* report detailing mirage effects on flagstone pavements. A paean to a sublime apparition, “real, but not true”. **Barbara Kiser**

Correspondence

Women's prize: be more generous

Your announcement of awards to celebrate women in science (*Nature* 556, 150; 2018) recalls another such prize announced more than 100 years ago (see *Science* 28, 832; 1908).

The Sarah Berliner Research Fellowship for Women came about in 1909 thanks to the efforts of mathematician Christine Ladd-Franklin, who completed all the requirements for a PhD at Johns Hopkins University in Baltimore, Maryland, in 1882 but did not receive her doctorate until 1926. (Similar to many US universities at the time, Johns Hopkins did not award PhDs to women.)

Ladd-Franklin convinced Emile Berliner — inventor of the gramophone, the flat-disc record and a type of microphone used in the first practical telephones — to endow a fellowship for female scientists in the name of his mother.

The fellowship enabled female scientists with a PhD to spend one year doing research at a US university. The stipend was US\$1,200 at a time when the average salary of “assistant professors in the leading universities” was \$1,800 (*Pop. Sci. Mon.* 76, 615; 1910). The fellowship, today administered by the American Association of University Women, is now worth \$30,000, matching the estimated 25-fold increase in prices since 1910.

It might be argued on these grounds that the *Nature* awards should be doubled from the stipulated total of \$15,200.

Richard P. Elinson *Huntington, New York, USA.*
elinson@duq.edu

Women's prize: act to boost all diversity

We applaud your announcement to further promote gender equality through your Inspiring Science and Innovating Science awards (*Nature* 556, 150; 2018). More such initiatives are sorely

needed to address the severe under-representation of people from minority racial and ethnic groups in science, technology, engineering, mathematics and medicine (STEMM).

These inequalities have arisen from much the same social and historical drivers that have made the contemporary STEMM sphere biased in favour of men (N. A. Fouad and M. C. Santana *J. Career Assess.* 25, 24–39; 2017). Barriers to equality — such as sexism, poverty and racism — compound one another, making it even harder for members of multiple minority groups to pursue an academic career (see *Nature* 547, 266–267; 2017).

Overcoming the planet's unprecedented challenges will demand all of our combined intellectual power — regardless of gender, race, ethnicity, sexuality, disability or any other diversity dimension that is currently under-represented in STEMM.

Ricardo Rocha, Fangyuan Hua *University of Cambridge, UK.*
rr552@cam.ac.uk

India's push for solar geoengineering

India has been contributing to the evaluation, discussion and implementation of solar-geoengineering research for almost a decade, in line with the call by A. Atiq Rahman and colleagues for developing countries to take the lead in this realm (see *Nature* 556, 22–24; 2018).

The Indian government's Department of Science and Technology launched a major research initiative in 2017 at the Indian Institute of Science in Bangalore to understand the implications of solar geoengineering on developing countries. The first annual meeting of experts and policymakers to discuss how this research could be done in India was held in 2017.

The department has also funded a geoengineering

climate-modelling research programme over the past five years. This has revealed, for example, how solar geoengineering could affect the global water cycle and extreme events and cyclones in the Bay of Bengal (see G. Bala and B. Nag *Clim. Dyn.* 39, 1527–1542; 2012; and A. Nalam *et al. Clim. Dyn.* 50, 3375–3395; 2018).

Furthermore, New Delhi's Council on Energy, Environment and Water has held three international conferences since 2011 to identify India's role in developing regional and global governance of solar-geoengineering research and technologies.

Govindasamy Bala *Indian Institute of Science, Bangalore, India.*

Akhilesh Gupta *Department of Science and Technology, Government of India, New Delhi, India.*
gbala@iisc.ac.in

Cooperate on research integrity

As an institutional research-integrity officer, I see first-hand how cooperation between journals and institutions is crucial for addressing research misconduct. As both camps consider utilizing the latest tools for detecting image duplication (see, for example, *Nature* 555, 18; 2018), it is important that they work closely together to deal with uncovered issues.

Last year's CLUE Recommendations put forward best practices for cooperation between editors and institutions to ensure research integrity and to protect the scientific record (E. Wager *et al.* Preprint at bioRxiv <https://doi.org/10.1101/139170>; 2017). For example, institutions need to be more willing to share information with journals, including research-misconduct reports. They should also consider asking journals to correct or retract publications as soon as data are known to be false, rather than waiting for

lengthy misconduct processes to be completed. And when research misconduct is suspected, journals should consider contacting institutions directly so that raw data can be properly secured.

Adopting best practices and cultivating strong partnerships are in everyone's best interests.

Lauran Qualkenbush *Northwestern University, Chicago, Illinois, USA.*
lhaney@northwestern.edu

Microbes set to alter the economy

Gene editing stands to accelerate the engineering of microbes for industrial production of food ingredients, pharmaceuticals, biofuels and biomaterials. There is a risk, however, that microbial biotechnologies could destabilize economies and employment in the developing world that depend on supplying naturally occurring ingredients. For example, a biosynthetic process for making a precursor of the antimalarial drug artemisinin has been developed, which could threaten the jobs of farmers who harvest its natural source, the plant *Artemisia annua*.

Microbial processes hold promise for global sustainable development: they are cheaper, consume less energy and pollute less than oil-based manufacturing, and they use renewable feedstocks (V. de Lorenzo *et al. EMBO Rep.* 19, e45658; 2018). Yet it is imperative that international stakeholders assess and address any social-justice problems that could arise from such applications (see C. G. Acevedo-Rocha in *Ambivalences of Creating Life* 9–53; Springer, 2016). Long-term commitment will be necessary to close the communication gap between scholars from different disciplines, cultures, values and generations (see also S. Jasanoff and J. B. Hurlbut *Nature* 555, 435–437; 2018).

Carlos G. Acevedo-Rocha *Biosyntia, Copenhagen, Denmark.*
car@biosyntia.com

Peter Grünberg

(1939–2018)

Physicist who revolutionized data storage with work on magnetism in nanomaterials.

Peter Grünberg was one of the first physicists to understand the potential of nascent nanotechnologies for fundamental research. He discovered giant magnetoresistance, or GMR: a large change in electrical resistance induced by a small magnetic field in stacks of ultrathin magnetic and non-magnetic layers. For this, he won a share of the 2007 Nobel Prize in Physics (as did I; we independently discovered the same effect). Ultimately, his work led to the development of hard-disk drives and greatly increased data storage. It also kicked off the field of spintronics. Grünberg died on 7 April 2018, aged 78.

Grünberg was born in 1939 in Pilsen, Bohemia, then a German protectorate, now part of the Czech Republic. In 1945, his family left for West Germany. There, at 19, Grünberg went to study physics at the Goethe University Frankfurt; he then did a PhD at the Technical University of Darmstadt.

For his PhD, he used optical spectroscopy to determine the energy levels of rare-earth ions in magnetic garnet crystals. In his postdoc, he turned another spectroscopy technique — Raman scattering — on garnets, this time at Carleton University in Ottawa, Canada, from 1969 to 1971. And in 1972, thanks to his expertise in the spectroscopic study of magnetic materials, Grünberg was offered a post at the newly founded Institute for Magnetism at the Jülich Research Centre in Germany.

Here, Grünberg quickly demonstrated his pioneering spirit, developing the spectroscopy technique of Brillouin light-scattering spectroscopy (BLS). BLS examines the inelastic scattering of light; it can probe both the ground state of magnetic materials and their excited states. In the 1970s, physicists were struggling to pick up the specific excitation modes expected to occur at the surface of magnetic materials. Grünberg singled out these modes, and identified them as spin waves of the Damon–Eshbach type.

During a sabbatical at Argonne National Laboratory in Illinois in 1985, Grünberg used an emerging technique of growing metals on single crystals to extend his BLS experiments to layers of magnetic materials less than 1 nanometre thick. This led to his first major discovery. In a sandwich of magnetic iron, non-magnetic chromium and more iron, he and his co-workers demonstrated the existence of antiferromagnetic exchange coupling between the iron layers across chromium (P. Grünberg *et al.* *Phys. Rev. Lett.* **57**, 2442; 1986). This was the



first demonstration of a quantum effect in magnetism. The coupling results from the interference between electronic wave functions reflected at the surface of the magnetic layers. For me, it was also the revelation of a nanostructure in which I could test some of my own ideas about magnetoresistance.

Soon after, in early 1988, our two teams independently discovered GMR. Grünberg's group showed it in their iron–chromium–iron sandwich; my team in France, in a stack of 40 layers of iron and chromium.

It was clear from the outset that GMR would have vast applications — especially because it could happen at room temperature. It could be used to detect small magnetic fields easily, which proved useful for magnetic sensors and, in particular, for reading magnetic hard disks. It was also the first example of electronics that exploit both the spin and charge of electrons — a field today called spintronics.

I met Grünberg in August 1988 at the International Colloquium on Magnetic Films and Surfaces in Le Creusot, a small town in the middle of France. We presented our respective results, and came to the conclusion that we had made the same discovery. To our delight, most participants seemed to think it was important: we celebrated with a couple of glasses of local wine, a Pommard burgundy. That night, we had the feeling that the conference concert, with pieces played on piano and violin by some of our colleagues, was in celebration of our work. Grünberg, a skilled guitarist, did not play that time, but I had several opportunities to hear him play.

Grünberg had an impressive vision for

smart GMR-based devices, including the spin-valve concept that he patented and was later developed at IBM for use in hard disks. The first commercial GMR-based hard disks appeared in 1997. Since then, their data-storage capacity has risen by almost three orders of magnitude. Grünberg conceived of many other devices, from magnetic sensors to the compass used in smartphones today.

The discovery of GMR kicked off an intense period of research activity, including experiments on GMR and interlayer exchange coupling in a great diversity of magnetic multilayers. At the same time, the theory behind GMR was developed. Grünberg's Jülich team took a semi-classical approach; I worked in collaboration with New York University on a quantum approach.

The burgeoning field of spintronics has yielded fascinating results. In 1995, Terunobu Miyazaki and Nobuki Tezuka in Japan, and a US group led by Jagadeesh Moodera, independently showed that quantum tunnelling of electrons between magnetic layers gives rise to much larger magnetoresistances than with GMR (T. Miyazaki and N. Tezuka *J. Magn. Magn. Mat.* **139**, L231–L234; 1995; J. S. Moodera *et al.* *Phys. Rev. Lett.* **74**, 3273; 1995). In 1996, John Slonczewski at IBM in Yorktown Heights, New York, and Luc Berger at Carnegie Mellon University in Pittsburgh, Pennsylvania, introduced the concept that transferring spins between magnetic materials could create torque on their magnetization direction (J. C. Slonczewski *J. Magn. Magn. Mat.* **159**, L1–L7; 1996; L. Berger *Phys. Rev. B* **54**, 9353; 1996). Today, this mechanism is exploited to write non-volatile magnetic memories. Grünberg and his team contributed much to these research fields.

Grünberg was warmly esteemed by colleagues around the world for his great creative talent in physics, and for his integrity and modesty. To me, he was also a good friend. And I liked his sense of humour. Having shared emotional and amusing moments in Stockholm, we often recalled our procession to the Nobel banquet, when neither of us was completely successful in our attempts not to tread on the trains of the Royal Princesses.

Peter was a great physicist, and a gentle and sincere person. The spintronics and nanomagnetism community will miss him sorely. ■

Albert Fert is a physicist at the University Paris-Sud in Orsay and at UMPH, a joint laboratory of CNRS and Company Thales. e-mail: albert.fert@cnrs-thales.fr

Safeguarding stem-cell fidelity

A complex that includes the protein HP1 binds to specific regulatory DNA sequences to promote local compaction of genomic regions and inhibit associated genes that drive differentiation of specific cell lineages. [SEE LETTER P.739](#)

KRISTOFFER N. JENSEN
& MATTHEW C. LORINCZ

Mutations in the gene that encodes activity-dependent neuroprotective protein (ADNP) cause the rare neurodevelopmental disorder Helsmoortel–Van der Aa syndrome, which is characterized by intellectual disability and autism spectrum disorder¹. On page 739, Ostapcuk *et al.*² present a detailed study of the function of this transcription factor. They show that ADNP forms part of a nuclear complex that plays an essential part in maintaining the fidelity with which pluripotent stem cells give rise to the three primary cell lineages in the human body.

DNA is intimately associated with many proteins, including transcription factors and the histones around which it is packaged. Together, this DNA–protein complex makes up chromatin. Gene-poor genomic regions, which are associated with low levels of transcription, adopt a condensed chromatin structure known as heterochromatin, whereas gene-rich regions adopt a relatively open structure called euchromatin. Although ADNP is associated with heterochromatin³, it has been unclear whether this interaction is relevant to the traits seen in ADNP-deficient mouse embryos — which include defects in the formation of the structure that gives rise to the brain and spinal cord, and aberrant expression of genes normally expressed in the extra-embryonic cells that support embryo development⁴. In addition, the spectrum of ADNP's genomic targets in euchromatin has remained undefined.

Ostapcuk *et al.* isolated the DNA sequences bound by ADNP in mouse embryonic stem (ES) cells — a type of pluripotent cell derived from mouse embryos. They identified about 15,000 genomic sites at which ADNP was bound, most of which lay within genes or in nearby regulatory sequences that control gene expression. The authors showed that the expression of many of the genes that were bound by ADNP in wild-type mouse ES cells was upregulated in cells genetically engineered to lack this factor, supporting a direct role for ADNP in transcriptional repression. A subset of these genes encode proteins that promote differentiation into specific extra-embryonic

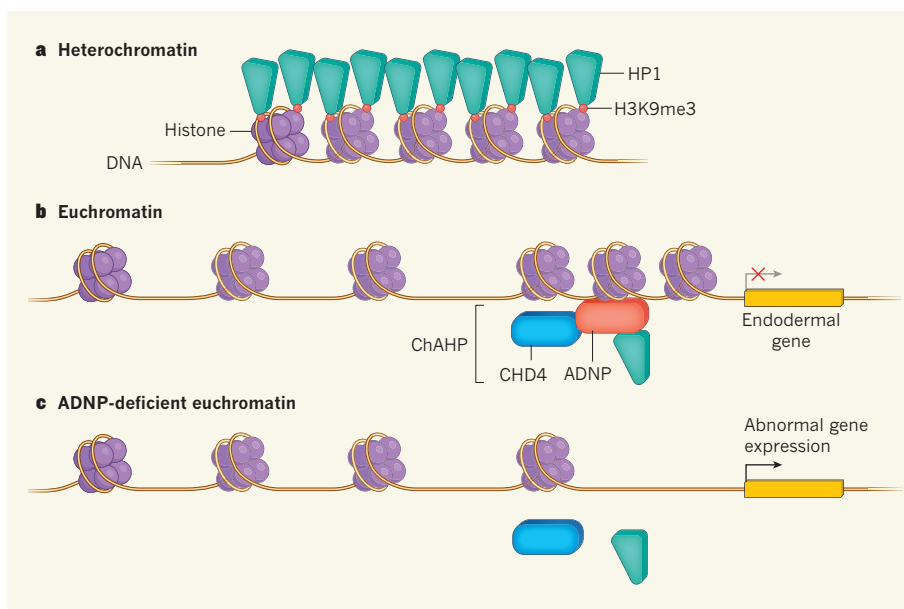


Figure 1 | Controlling gene expression through chromatin compaction. DNA is packaged around histone proteins in a complex called chromatin. **a**, HP1 proteins bind to methyl groups on histone H3 (a chemical modification called H3K9me3) to promote large-scale condensation of DNA into heterochromatin — a structure associated with transcriptional repression. **b**, Ostapcuk *et al.*² report another role for HP1 — interacting with the DNA-binding protein ADNP and the chromatin-remodelling protein CHD4 in a complex called ChAHP. This complex promotes local condensation of chromatin independent of H3K9me3, in regions of otherwise loosely packaged chromatin called euchromatin. Such binding inhibits the expression of associated genes — including those that promote stem-cell differentiation into cells of the ‘endodermal’ lineage. **c**, In cells devoid of ADNP, the ChAHP complex is disrupted and CHD4 and HP1 are not recruited, resulting in locally decondensed chromatin, abnormal gene expression and spontaneous stem-cell differentiation.

or embryonic tissues, in particular of the endodermal lineage (one of the primary cell lineages, which gives rise to the digestive and respiratory tracts). And, whereas wild-type mouse ES cells cultured under conditions that promote neuronal differentiation gradually became more neuron-like, ADNP-deficient cells failed to do so, and showed aberrant expression of endodermal genes.

Next, to determine whether ADNP mediates transcriptional silencing by recruiting corepressors, Ostapcuk and colleagues used an unbiased approach to identify proteins that interact with ADNP. Their screening yielded two types of chromatin-associated protein. The first, CHD4, is a chromatin-remodelling protein previously implicated in the control of genes associated with pluripotency and differentiation⁵. The second, HP1, has a role in transcriptional repression and packaging of

heterochromatin. Together, ADNP, CHD4 and HP1 form a stable complex that the authors dubbed ChAHP.

ADNP probably interacts with HP1 proteins through a well-documented HP1-binding domain in ADNP called a PXVXL motif⁴. There are three HP1 isoforms (HP1 α , HP1 β and HP1 γ), but Ostapcuk *et al.* found that only HP1 γ and, to a lesser extent, HP1 β interacted with ADNP, consistent with a report published earlier this year⁶. Furthermore, the authors showed that most genomic regulatory regions bound by ADNP in wild-type cells were also bound by HP1 γ and HP1 β . By contrast, another study showed that ADNP does interact with HP1 α in embryonic cancer cells⁴. Reconciling these apparently contradictory results, Ostapcuk *et al.* found that ADNP can interact with HP1 α , but only in the absence of HP1 β and HP1 γ . When the authors

examined cells lacking different combinations of HP1 isoforms, they found that all three had to be deleted to mimic the effects of ADNP mutation on gene expression. This indicates that the three HP1 proteins are functionally redundant — that is, they can compensate for one another in ChAHP.

Why might ADNP preferentially interact with HP1 γ and HP1 β ? HP1 α has an amino-terminal region, unique among HP1 proteins, through which it promotes the self-organization of heterochromatin into liquid-like droplets⁷. This process, called phase separation, probably minimizes interactions with other nuclear proteins, maintaining the condensed state of heterochromatin. However, ChAHP-mediated silencing of euchromatic genes must be reversible to enable genes to respond to differentiation cues during development, perhaps making phase separation less desirable at these genomic regions.

HP1 proteins interact with PXVXL-containing factors such as ADNP through a carboxy-terminal 'chromoshadow' domain. But in heterochromatin, HP1 binds histone H3 through an evolutionarily conserved amino-terminal chromodomain⁸, at sites where the histone is tagged by methyl groups on amino-acid residue lysine 9 — a chemical modification dubbed H3K9me3. It is through this interaction that HP1 proteins induce chromatin condensation. Might H3K9me3 also promote HP1 binding at ChAHP-bound regions? In favour of this hypothesis, HP1 can recruit ADNP to H3K9me3-marked heterochromatin³. However, Ostapcuk and colleagues show that ChAHP can efficiently bind to its DNA targets even if HP1 is engineered to lack its chromodomain. Furthermore, regulatory sequences bound by ChAHP lacked H3K9me3. Therefore, H3K9me3 is unlikely to have a role in ChAHP-mediated transcriptional silencing.

People with Helsmoortel–Van der Aa syndrome generally have ADNP mutations that produce a truncated protein lacking the DNA-binding domain and the PXVXL motif⁹. To investigate whether this mutation disrupts ChAHP-complex formation, Ostapcuk *et al.* expressed one such patient-derived mutant protein in mouse ES cells. The mutant ADNP failed to bind to HP1 β or HP1 γ , and genes normally bound by ChAHP were aberrantly expressed. Furthermore, analysis of chromatin accessibility in ADNP-deficient ES cells revealed that the mutation led to opening of chromatin in regions immediately flanking ADNP-binding sites.

Together, Ostapcuk and colleagues' findings demonstrate that ChAHP does not generate broad swathes of heterochromatin, as observed at H3K9me3-marked regions bound by HP1. Instead, the complex generates focused regions of condensed chromatin that inhibit the transcription of differentiation-promoting genes. Aberrant expression of such genes in the absence of the ChAHP complex

is probably a crucial factor in the aetiology of Helsmoortel–Van der Aa syndrome (Fig. 1).

Although Ostapcuk and co-workers' study focused on the interplay between ADNP, CHD4 and HP1, the researchers found many fewer genes upregulated in mouse ES cells lacking ADNP than in those lacking all three HP1 isoforms. And, consistent with previous reports^{10,11}, the authors' analysis of HP1-interacting proteins revealed a plethora of overlapping and isoform-specific binding partners, many of which have DNA-binding activity. Notably, mutations in several of these HP1-interacting transcription factors are implicated in other rare syndromes associated with intellectual disability, including in genes that encode the proteins AHDC1 (ref. 12), CHAMP1 (ref. 13) and POGZ (ref. 14). Thus, it is tempting to speculate that HP1 proteins act as co-repressors for many as-yet-undescribed DNA-binding complexes that regulate the expression of distinct gene sets.

Ostapcuk and co-workers' study has revealed a key mechanism of HP1 recruitment to chromatin. Their work sets the stage for future studies on the broader role of this enigmatic co-repressor in gene regulation and local chromatin compaction. ■

QUANTUM PHYSICS

Molecular dynamics simulated by photons

The microscopic behaviour of molecules can be difficult to model using ordinary computers because it is governed by quantum physics. A photonic chip provides a versatile platform for simulating such behaviour. [SEE ARTICLE P.660](#)

FABIEN GATTI

Quantum-computing devices could one day outperform ordinary computers, particularly in the simulation of quantum systems. Such devices share their quantum nature with the system to be simulated and are therefore inherently suited to describing quantum phenomena¹. On page 660, Sparrow *et al.*² report a device based on a single photonic chip that can simulate a range of quantum dynamics associated with different molecules. The results are in excellent agreement with simulations carried out by ordinary computers, reaffirming the potential of quantum technology in this area.

In conventional industrial chemistry, the yields of chemical processes are optimized by controlling macroscopic variables, such as temperature and pressure. But the use of high temperatures and pressures wastes a substantial amount of energy and generates unwanted by-products, leading to high energy consumption and pollution. To overcome these issues,

Kristoffer N. Jensen and Matthew C. Lorincz are in the Department of Medical Genetics, Life Sciences Institute, University of British Columbia, Vancouver, British Columbia V6T 1Z3, Canada.
e-mail: matthew.lorincz@ubc.ca

1. Helsmoortel, C. *et al.* *Nature Genet.* **46**, 380–384 (2014).
2. Ostapcuk, V. *et al.* *Nature* **557**, 739–743 (2018).
3. Mosch, K., Franz, H., Soeroes, S., Singh, P. B. & Fischle, W. *PLoS ONE* **6**, e15894 (2011).
4. Mandel, S., Rechavi, G. & Gozes, I. *Dev. Biol.* **303**, 814–824 (2007).
5. Zhao, H. *et al.* *J. Biol. Chem.* **292**, 8507–8519 (2017).
6. Zaidan, N. Z. *et al.* *Stem Cell Rep.* **10**, 627–641 (2018).
7. Larson, A. G. *et al.* *Nature* **547**, 236–240 (2017).
8. Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. *Nature* **410**, 116–120 (2001).
9. Gozes, I., Yeheskel, A., & Pasmanik-Chor, M. *J. Alzheimer's Dis.* **45**, 57–73 (2015).
10. Hauri, S. *et al.* *Cell Rep.* **17**, 583–595 (2016).
11. Nozawa, R.-S. *et al.* *Nature Cell Biol.* **12**, 719–727 (2010).
12. Xia, F. *et al.* *Am. J. Hum. Genet.* **94**, 784–789 (2014).
13. Tanaka, A. J. *et al.* *Cold Spring Harb. Mol. Case Stud.* **2**, a000661 (2016).
14. Ye, Y. *et al.* *Cold Spring Harb. Mol. Case Stud.* **1**, a000455 (2015).

This article was published online on 23 May 2018.

a promising optimization approach exploits the quantum nature of the reacting molecules.

A central tenet of quantum physics is the superposition principle, which asserts that possible quantum states of a system can be added together and the result will be another possible state. The non-classical aspect of this principle is demonstrated, for example, by quantum bits. These objects can exist in both an on state and an off state at the same time. Such states exhibit quantum coherence, which means that they are correlated in a non-classical way.

The ability to systematically control quantum coherence is considered one of the main challenges in energy science. Such control might enable the synthesis of highly desirable materials and devices, including superfluids (fluids that flow without resistance) and quantum computers. It could also give rise to more-efficient chemical processes than are currently possible.

In conventional chemistry, the quantum states involved in chemical processes are

incoherent. However, coherent superpositions of molecular states can be produced using the light emitted by a laser. The ability to consistently generate these superpositions could improve the efficiency of the corresponding chemical processes and reduce the energy required to control such processes. It might even open up chemical-reaction mechanisms that are otherwise inaccessible^{3,4}.

Laser pulses are the main tool for manipulating molecules in this field. Improvements in the design of these pulses, such as increases in power and tunability, as well as the ability to reduce the duration of the pulses to attosecond (10^{-18} s) timescales, have enabled greater control of light-induced processes in molecules⁵. Since the pioneering work of Ahmed Zewail, who was awarded the 1999 Nobel Prize in Chemistry (see go.nature.com/2idzowq), laser pulses have been used to study quantum coherence in chemistry⁶.

For example, quantum coherence has been used to enhance the rates of chemical reactions in biological systems at room temperature⁷. Such studies conclusively showed that quantum coherence can be partially preserved even in molecular systems open to the external environment⁸.

These experimental advances call for accurate models of the quantum evolution of molecular systems. This is a challenging task for quantum-computing devices, although much progress has been made, thanks to the development of improved algorithms for simulating quantum dynamics^{9,10}.

Sparrow and colleagues engineered a quantum-computing device that is based on a single photonic chip. They used the quantum superposition of photons in the chip to carry quantum information and to model molecular systems. By adjusting the optical circuitry of the chip, the authors simulated a range of quantum dynamics associated with different molecules.

The authors began by simulating vibrational excitations in a variety of four-atom molecules. They then modelled energy transport in the chemical bond of a protein and the transfer of vibrational energy in liquid water. Finally, they tested an algorithm designed to identify quantum states that can lead to the break-up of ammonia. The results of these simulations were in almost perfect agreement with those obtained using ordinary computers.

The first quantum revolution occurred at the turn of the twentieth century, and provided us with the physical laws that govern reality. Sparrow and colleagues have now simulated the time evolution of a quantum superposition of molecular states with the aid of an experimental device that uses the quantum superposition of photons. Such a feat suggests that we could be entering a second quantum revolution, in which the physical laws of nature are used to develop innovative technologies.

Despite these promising prospects, it is not difficult to envisage the problems that follow-up

studies will encounter. In this seminal work, the authors used rather simple molecular models, involving a limited number of mathematical terms. However, this number will increase exponentially when aiming to closely reproduce experimental conditions. Such an increase might dramatically enhance what the authors refer to as the “fundamental errors” in photonics, which include the loss of photons and the loss of quantum coherence.

Nevertheless, Sparrow and colleagues have demonstrated that simulations carried out by quantum-computing devices can be both reliable and efficient, by tackling problems that can be solved using well-established standard techniques. As the authors point out, slight improvements in their method could yield simulations that cannot be achieved using ordinary computers. ■

IMMUNOLOGY

Gut molecules control brain inflammation

Metabolite molecules produced by the gut's microbes activate immune cells in the brain called microglia, which signal to astrocyte cells to mediate responses to inflammation in the central nervous system. [SEE LETTER P.724](#)

HARTMUT WEKERLE

Some immunologists regard the central nervous system (CNS) as a no-man's-land, avoided by immune cells and therefore uninteresting. But, in fact, the CNS has a vigorous immune potential that remains dormant in normal conditions but is awakened after injury. The switch that controls the brain's immune microenvironment involves non-neuronal cells called glia — not only microglia, which are sometimes called the immune cells of the CNS, but also multifunctional cells called astrocytes¹. Rothhammer *et al.*² describe on page 724 how these two glial cell types communicate on a molecular level to influence inflammation in the CNS, and show that this interaction is controlled remotely by microbes that inhabit the gut.

A decade ago, the group that performed the current study, along with another research group, discovered^{3,4} an unexpected immunoregulatory role for a ligand-activated transcription factor called the aryl hydrocarbon receptor (AHR), which at the time was best known as a receptor for environmental toxins⁵. The two groups showed that AHR modulates the progression of experimental autoimmune encephalomyelitis (EAE) — an autoimmune disease in mice in which the immune system becomes overactive and attacks the CNS. EAE is often used a model

Fabien Gatti is at the Institut des Sciences Moléculaires d'Orsay, CNRS, Université de Paris-Sud, F-91405 Orsay, France.
e-mail: fabien.gatti@u-psud.fr

1. Feynman, R. P. *Int. J. Theor. Phys.* **21**, 467–488 (1982).
2. Sparrow, C. *et al. Nature* **557**, 660–667 (2018).
3. Moore, K. & Rabitz, H. *Nature Chem.* **4**, 72–73 (2012).
4. Brif, C., Chakrabarti, R. & Rabitz, H. *N. J. Phys.* **12**, 075008 (2010).
5. Corkum, P. B. & Krausz, F. *Nature Phys.* **3**, 381–387 (2007).
6. Gatti, F., Lasorne, B., Meyer, H.-D. & Nauts, A. *Applications of Quantum Dynamics in Chemistry* (Springer, 2017).
7. Collini, E. *et al. Nature* **463**, 644–647 (2010).
8. Scholes, G. D. *et al. Nature* **543**, 647–656 (2017).
9. Meyer, H.-D., Manthe, U. & Cederbaum, L. S. *Chem. Phys. Lett.* **165**, 73–78 (1990).
10. Meyer, H.-D., Gatti, F. & Worth, G. A. *Multidimensional Quantum Dynamics: MCTDH Theory and Applications* (Wiley-VCH, 2009).

of multiple sclerosis (MS). Initially, the groups focused on how AHR might affect EAE by regulating pathogenic and protective subsets of immune cells outside the CNS. But it later emerged that AHR is also strongly expressed in the CNS, particularly in microglia and astrocytes⁶, raising the question of whether AHR in the CNS has a role in autoimmune diseases.

In the current study, Rothhammer *et al.* induced EAE in mice that had been genetically engineered so that AHR could be deleted in microglia (but not in other brain cells or immune cells) by a drug treatment. Elimination of microglial AHR substantially exacerbated EAE in the AHR-depleted mice, but left immune responses outside the CNS unaltered. This finding suggests that AHR activation in microglia inhibits inflammation in the CNS.

Microglia rarely act alone. Instead, they often team up with other cell types to respond to the stimuli that activate them. For example, after being activated, microglia can instruct certain astrocytes to attack local neurons⁷. Rothhammer and colleagues found that AHR-deficient microglia activated by EAE triggered exaggerated inflammatory responses in local astrocytes. Next, the authors used bioinformatics to analyse the gene-expression pathways altered in these glia. This analysis suggested that unexpected proteins signal from microglia to astrocytes.

The usual suspects in such cases are

pro-inflammatory signalling molecules, but Rothhammer *et al.* showed that AHR in microglia directly regulates the expression of genes that encode the proteins TGF- α and VEGF-B (Fig. 1) — neither of which has previously received much attention from neuroimmunologists. Subsequent detailed *in vitro* and *in vivo* analyses confirmed that TGF- α and VEGF-B regulate the pro-inflammatory reactivity of astrocytes. TGF- α dampens astrocyte inflammatory responses to EAE, and its expression in microglia is inhibited by AHR deletion. Conversely, VEGF-B enhances responses to EAE, and its expression is promoted by AHR deletion.

So, the activity of microglia and astrocytes is modulated by AHR during brain inflammation in autoimmune disease. But which signals might modulate microglial AHR? In addition to environmental toxins, AHR is bound by a broad range of molecules, including dietary derivatives⁸. In particular, food plants such as broccoli and other members of the cabbage family contain components that bind AHR either directly or after being processed into metabolite molecules, such as derivatives of tryptophan (Trp), by gut microbes⁹. Rothhammer *et al.* fed their mice diets either depleted or enriched in Trp. Trp depletion exacerbated EAE in wild-type mice, whereas enrichment ameliorated the effects of the disease. By contrast, neither diet had any effect on the progress of EAE in AHR-deficient animals, as might have been predicted — in these animals, Trp cannot bind to AHR to dampen immune responses.

To determine whether their work is likely to have implications for humans, the authors verified basic elements of their analyses in tissue samples from people with MS, in which an autoimmune attack drives glial inflammation, destruction of nerve processes and their insulating myelin sheaths, and ultimately scar formation¹⁰. The group found that AHR, TGF- α and VEGF-B were expressed in microglia-like cells in MS tissues. Levels of the proteins were higher in newly inflamed regions than in old scar tissue or unaffected surrounding tissue. This suggests (but does not prove) that TGF- α and VEGF-B have a role in the formation of MS scar tissue.

Rothhammer and colleagues' work sheds light on the complex regulation of inflammatory reactivity in the CNS and adds another facet to our understanding of the gut–brain connection. Robust regulation of inflammatory responsiveness is essential for proper CNS function. Deficient regulation, with unrestrained inflammatory episodes, leads to sickness, irreversible cell loss and scar formation¹¹, whereas compromised inflammatory reactivity can result in tumour formation and opportunistic infection¹². The authors' findings are therefore likely to have implications beyond MS.

The interactions between the gut, microglia and astrocytes outlined by Rothhammer *et al.*

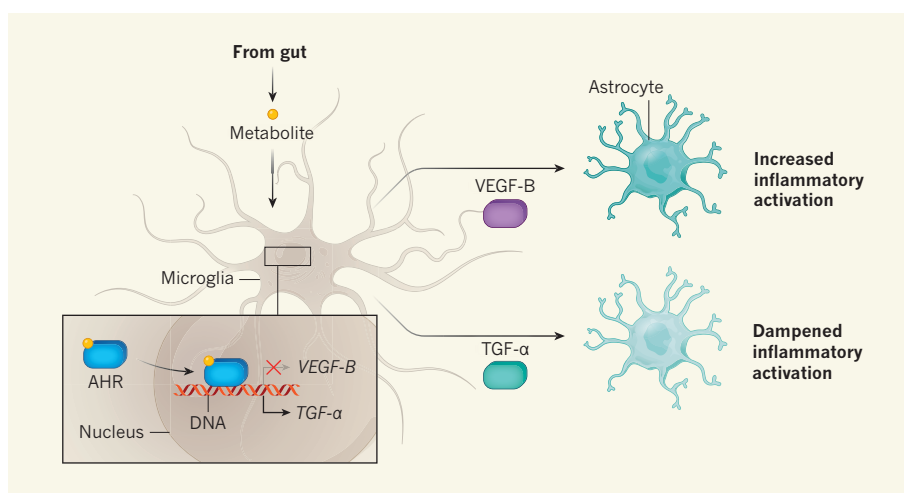


Figure 1 | Long-distance regulation of immune cells in the brain. Gut bacteria process the dietary component tryptophan to produce metabolite molecules that enter the central nervous system (CNS). In the brain, these metabolites act as ligands for the aryl hydrocarbon receptor (AHR) — a transcription factor expressed in cells called microglia and astrocytes that mediate responses to inflammation in the CNS. Rothhammer *et al.*² report that, when activated in microglia, AHR binds to the genes that encode the proteins VEGF-B and TGF- α , inhibiting the expression of the former and promoting that of the latter. Any VEGF-B released from microglia enhances the responsiveness of astrocytes to CNS inflammation. By contrast, TGF- α dampens astrocyte responsiveness.

are not the only mechanisms that safeguard inflammatory responses in the brain¹³. It will be of interest to examine how other regulators of the CNS immune microenvironment modulate the newly identified signalling pathway. These factors include the cells associated with the cerebral blood vessels, as well as active neurons. Indeed, pharmacological silencing of neurons leads to the activation of neighbouring microglia¹⁴.

That the behaviour of microglia can be controlled remotely by intestinal products is intriguing, although not without precedent. A flurry of observations previously linked the CNS to the gut and its microbial contents. Neuronal pathways, hormones, microbial molecules and metabolites are all involved in signalling between these regions¹⁵. Specifically, short-chain fatty acids produced by gut bacteria can modulate microglia cells¹⁶, and tryptophan metabolites act directly on astrocytes⁶. Nonetheless, the current findings broaden our understanding of the gut–brain connection. The authors speculate that this pathway might support the repair of injured neural cells.

As Rothhammer and colleagues point out, their experimental observations might lead to new therapeutic approaches to quelling unwanted CNS inflammation, and possibly to supporting neuronal repair. First, enhancement of TGF- α and blockade of VEGF-B might reduce CNS inflammation to an acceptable, non-toxic level. Second, clinical CNS

inflammation could be dampened indirectly by means of the gut. Dietary protocols that promote anti-inflammatory regulation could be a promising non-invasive approach to treating brain inflammation. It is to be hoped that diets that have been proposed as effective medications for diseases such as MS, but whose effectiveness has yet to be formally proved¹⁷, will now be re-examined. ■

Hartmut Wekerle is at the Max Planck Institute of Neurobiology, and at the Institute of Clinical Neuroimmunology, LMU Munich, D-82152 Martinsried, Germany.
e-mail: hwekerle@neuro.mpg.de

1. Salter, M. W. & Stevens, B. *Nature Med.* **23**, 1018–1027 (2017).
2. Rothhammer, V. *et al.* *Nature* **557**, 724–728 (2018).
3. Quintana, F. J. *et al.* *Nature* **453**, 65–71 (2008).
4. Veldhoen, M. *et al.* *Nature* **453**, 106–109 (2008).
5. Okey, A. B., Riddick, D. S. & Harper, P. A. *Trends Pharmacol. Sci.* **15**, 226–232 (1994).
6. Rothhammer, V. *et al.* *Nature Med.* **22**, 586–597 (2016).
7. Liddelow, S. A. *et al.* *Nature* **541**, 481–487 (2017).
8. Stockinger, B., Di Meglio, P., Gialitakis, M. & Duarte, J. H. *Annu. Rev. Immunol.* **32**, 403–432 (2014).
9. Veldhoen, M. & Brucklacher-Waldert, V. *Nature Rev. Immunol.* **12**, 696–708 (2012).
10. Lassmann, H. & Wekerle, H. in *McAlpine's Multiple Sclerosis* 4th edn (eds Compston, A. *et al.*) 557–600 (Churchill Livingstone, 2006).
11. Sofroniew, M. V. *Neuroscientist* **20**, 160–172 (2014).
12. Monaco, M. C. G. & Major, E. O. *Front. Immunol.* **6**, 159 (2015).
13. Li, Q. & Barres, B. A. *Nature Rev. Immunol.* **18**, 225–242 (2018).
14. Neumann, H. *Glia* **36**, 191–199 (2001).
15. Sampson, T. R. & Mazmanian, S. K. *Cell Host Microbe* **17**, 565–576 (2015).
16. Erny, D. *et al.* *Nature Neurosci.* **18**, 965–977 (2015).
17. Riemann-Lorenz, K. *et al.* *PLoS ONE* **11**, e0165246 (2016).

This article was published online on 16 May 2018.

Enzyme illuminates bacterial ubiquitination

Structural analysis reveals how a bacterial enzyme catalyses attachment of the protein tag ubiquitin to host proteins, illuminating a process that allows pathogenic bacteria to subvert host-cell function. [SEE ARTICLE P.674 & LETTERS P.729 & P.734](#)

KATHY WONG & KALLE GEHRING

Ubiquitination is a type of protein modification in which the protein ubiquitin is attached to a target protein. In eukaryotes (organisms that include fungi, plants and animals), the addition of a ubiquitin tag can act as a signal for various cellular processes. A prime example is the destruction of ubiquitinated proteins by a eukaryotic protein complex called the proteasome. The ubiquitination process is also the target of many bacterial pathogens, which have developed techniques to hijack it for their own benefit. In papers in *Nature*, Akturk *et al.*¹ (page 729), Dong *et al.*² (page 674) and Kalayil *et al.*³ (page 734) describe the X-ray crystal structure of the bacterial enzyme SdeA, which catalyses ubiquitination. And, writing in *Cell*, Wang *et al.*⁴ report the structure of a bacterial enzyme called SidE from the same protein family as SdeA.

The eukaryotic ubiquitination pathway requires a three-enzyme cascade⁵. An enzyme called E1 activates ubiquitin using a molecule of ATP and a magnesium ion (Mg^{2+}) to covalently bind the ubiquitin through a type of linkage called a thioester bond. The activated ubiquitin is then transferred to downstream enzymes, which attach ubiquitin through

an isopeptide bond to a lysine amino-acid residue in the target protein. The discovery⁶ of the SidE family of ubiquitin ligase enzymes in the bacterial pathogen *Legionella pneumophila* revealed a ubiquitination pathway with striking differences from the eukaryotic system. Not only can SidE ligases carry out the complete process without the aid of other enzymes, but this pathway also generates a different form of ubiquitin, termed phosphoribosylated ubiquitin (PR-Ub), in which a phosphoribose-sugar linkage attaches ubiquitin to the target protein⁷.

The bacterial ubiquitination pathway requires a molecule of NAD^+ instead of the ATP and Mg^{2+} used by eukaryotes. In the first step, the mono-ADP-ribosyltransferase (mART) domain of SdeA uses ubiquitin and NAD^+ to covalently attach an adenosine diphosphate ribose (ADPR) molecule to an arginine residue (Arg42) of ubiquitin⁶, producing an ADPR-Ub molecule. The phosphodiesterase (PDE) domain of SdeA then cleaves this ADPR-Ub to release the molecule AMP, generating PR-Ub, and this group forms a bond with a serine residue in the target protein⁷. However, the molecular details of how ubiquitination is catalysed were a mystery until now.

The four papers^{1–4} provide a detailed picture of the bacterial reaction pathway, with

complementary insights into the catalytic mechanisms. Akturk, Dong and Kalayil, and their respective colleagues, report atomic structures of SdeA's catalytic core, which consists of the PDE and mART domains.

Dong *et al.* imaged the largest fragment of SdeA, which includes part of the protein's carboxy-terminal domain. They observed that the C-terminal domain is required to anchor the PDE and mART domains, stabilizing the enzyme in an active conformation. The structure of SidE reported by Wang and colleagues reveals that its catalytic domains are similar to those of SdeA, but the authors conclude that the C-terminal domain mediates SidE dimerization. This disparity might result from differences in the experimental conditions used by the two groups, or might reflect specialized functions of the individual SidE family members.

The generation of ADPR-Ub from ubiquitin and NAD^+ in the first step of the reaction is revealed in a structure presented by Dong *et al.* of the mART domain in complex with ubiquitin and the molecule NADH, which is similar to NAD^+ but can inhibit catalysis by the enzyme. This revealed that the Arg42 residue in ubiquitin that becomes modified is located too far away from the ribose group of NADH for modification to occur directly. By contrast, another of ubiquitin's arginine residues, Arg72, which was previously shown to be important in SdeA-mediated ubiquitination⁷, is located much closer to the enzyme-bound NADH. The authors used computer simulations of the complex, called molecular dynamics, to show that Arg72 and one other arginine residue (Arg74) anchor ubiquitin to mART. Once the nicotinamide group from NADH is released from the enzyme, a conformational change can occur, allowing Arg42 to replace Arg72 in the active site. This model explains why ADPR attaches selectively to Arg42 and not to

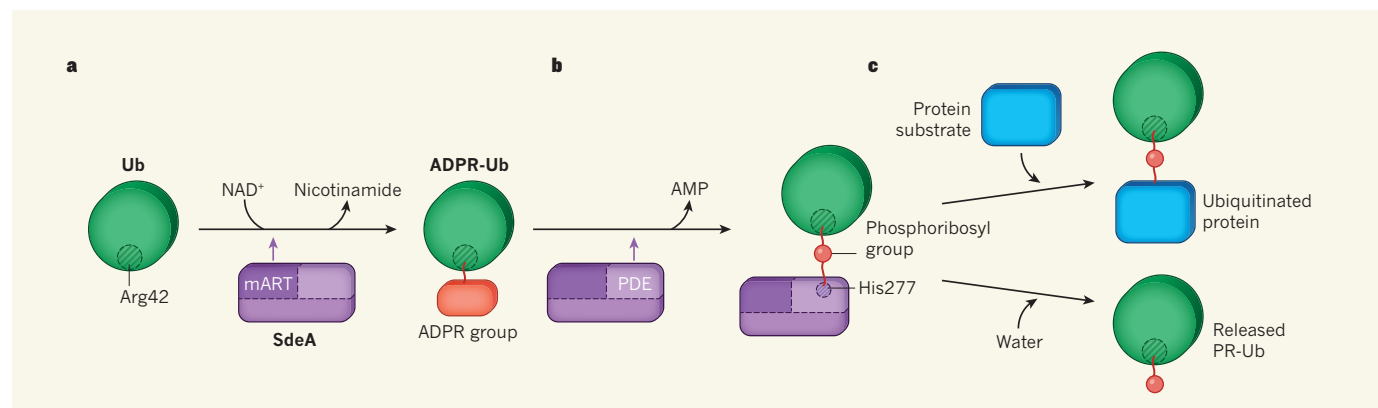


Figure 1 | The ubiquitination mechanism used by bacteria. Four papers^{1–4} provide complementary insights into how bacterial enzymes from the SidE family mediate the process in which the protein ubiquitin (Ub) is attached to a target protein. Akturk *et al.*¹, Dong *et al.*² and Kalayil *et al.*³ report the structure of the enzyme SdeA, and Wang *et al.*⁴ present the structure of the enzyme SidE. **a**, In the first step, the enzyme's mART domain processes NAD^+ and adds an adenosine diphosphate ribose (ADPR) group to the amino-acid residue arginine 42 (Arg42) of ubiquitin. This generates ADPR-Ub in a reaction that releases nicotinamide. Dong *et al.*²

reveal that Arg72 of ubiquitin (not shown) helps to anchor ubiquitin to the enzyme. **b**, ADPR-Ub is then processed by the enzyme's PDE domain. The molecule AMP is released in a reaction that generates ubiquitin bound to a phosphoribosyl group (PR-Ub); the phosphoribosyl group, in turn, is covalently attached to the enzyme's amino-acid residue histidine 277 (His277). **c**, If a protein substrate enters the enzyme's active site, the enzyme catalyses the attachment of PR-Ub to a serine residue (not shown) on the protein substrate. If, instead, water enters the active site, PR-Ub is released.

other arginines in ubiquitin, but further study is warranted to fully understand the process.

As the reaction proceeds, ADPR-Ub is processed by the PDE domain and PR-Ub is attached to a serine residue on a substrate protein. In addition to their studies of SdeA, Akturk *et al.* present the structure of ADPR-Ub in complex with SdeD, a member of the SidE family that contains only a PDE domain. Kalayil *et al.* used mass spectrometry techniques to study the SdeA catalytic intermediates at this stage. Both groups propose a two-step reaction mechanism for SdeA on the basis of studies of SdeA or SdeD.

First, the Glu340 amino-acid residue of SdeA binds ADPR-Ub. The His277 residue of SdeA interacts with a phosphate group on ADPR-Ub, resulting in the release of a molecule of AMP. Second, His407 activates the hydroxyl group of a serine residue on the target protein, which enables the attachment of PR-Ub to the serine. Using a mutated version of SdeA in which the histidine residue at position 407 was replaced with asparagine to trap a catalytic intermediate, Kalayil *et al.* captured PR-Ub bound to His277 of SdeA (Fig. 1), confirming the catalytic mechanism. Wang *et al.* report the structures of related complexes of ADPR and ubiquitin with SidE.

If a water molecule enters the PDE domain's active site instead of a serine amino-acid residue, the reaction product released is unbound PR-Ub. PR-Ub can inhibit host E1-dependent ubiquitination because the PR modification prevents this form of ubiquitin from being a substrate for eukaryotic ubiquitination enzymes⁷. Kalayil *et al.* answered the question of whether the pathogenicity associated with SdeA arises from the generation of unbound PR-Ub or from the ubiquitination of host proteins. The authors tested bacterial mutants lacking SidE proteins that were engineered to express either wild-type SdeA or a mutant version of SdeA that generates only unbound PR-Ub. The authors observed that the bacteria that express mutant SdeA were unable to grow in host cells, indicating that the enzyme's key role is ubiquitination of host proteins.

The role of PR-Ub is an emerging topic in the field of ubiquitin research. These structures of SidE family members now pave the way for more questions to be answered. For example, how is ADPR-Ub shuffled between the PDE and mART domains? The active sites of the PDE and mART domains are far apart (55 Å) and do not face each other. There is conflicting evidence as to whether SidE proteins exist as monomers or dimers, and, as a result, there are different models of how the gap between the domains might be bridged.

And what range of functions does the enzyme's C-terminal domain have? The C-terminal domain stabilizes the catalytic core in SdeA but mediates protein dimerization in SidE. Dong *et al.* observed that ubiquitin molecules bind to the C-terminal domain of SdeA and induce a large conformational change

in the enzyme, which suggests a possible regulatory role for this domain.

How many host proteins are ubiquitinated by SidE-family ligases? So far, only a few SdeA substrates have been identified^{6,8,9}; these include the GTPase enzymes Rab and Rag, as well as the protein RTN4. From analysis of the ubiquitination sites in host proteins, Kalayil *et al.* and Wang *et al.* propose that the ligase enzyme specifically targets serine residues in disordered protein regions.

Finally, perhaps the most exciting question still to be answered is this: do enzymes that mediate this type of ubiquitination process also exist in eukaryotes? ■

QUANTUM MATERIALS

Spinning on the edge of graphene

Long-sought evidence has been found of magnetism at the edges of graphene, a two-dimensional form of carbon. The findings might enable the development of the logic gates needed for quantum computers. SEE LETTER P.691

FERNANDO LUIS & EUGENIO CORONADO

The 2D form of carbon known as graphene has many potentially useful properties, but is usually not magnetic when pristine. However, theoretical predictions suggest that the edges of graphene sheets should become magnetic when they have a zigzag arrangement of carbon atoms¹. Observing this effect has been challenging because of the difficulties of detecting the predicted minute magnetic signal and because it is hard to fabricate defect-free edges that have the required shape. On page 691, Slota *et al.*² report a method for making nanometre-wide graphene ribbons in solution, and thereby for producing nanoribbons with well-defined zigzag edges 'decorated' with organic radical molecules that bear electron spins — a quantum property of electrons that is associated with magnetism. The authors' results provide solid evidence of magnetism at graphene edges, and show that edge spins have potentially useful quantum dynamics.

Magnetic forms of graphene would be useful for spintronics, a technology that forms the basis of today's magnetic data storage^{3,4}. But the main interest in generating magnetic edge states in graphene is for quantum technologies. Electron spins can adopt two orientations relative to an external magnetic field, and these could be used to encode the '0' and '1' states of a quantum bit (qubit), the basic information unit of future quantum computers and quantum-simulation devices.

The quantum states of a qubit must be

Kathy Wong and Kalle Gehring are in the Department of Biochemistry, McGill University, Montreal, Quebec H3G 0B1, Canada. e-mail: kalle.gehring@mcgill.ca

1. Akturk, A. *et al.* *Nature* **557**, 729–733 (2018).
2. Dong, Y. *et al.* *Nature* **557**, 674–678 (2018).
3. Kalayil, S. *et al.* *Nature* **557**, 734–739 (2018).
4. Wang, Y. *et al.* *Cell* **173**, 1231–1243 (2018).
5. Pickart, C. M. & Eddins, M. J. *Biochim. Biophys. Acta* **1695**, 55–72 (2004).
6. Qiu, J. *et al.* *Nature* **533**, 120–124 (2016).
7. Bhogaraju, S. *et al.* *Cell* **167**, 1636–1649 (2016).
8. Kotewicz, K. M. *et al.* *Cell Host Microbe* **21**, 169–181 (2017).
9. De Leon, J. A. *et al.* *Cell Rep.* **21**, 2031–2038 (2017).

strongly coupled to external control stimuli that drive the qubit's operation, but they must also be isolated from random external perturbations that can irreversibly upset the 'coherent' evolution of such quantum states (coherence is the existence of non-classical correlations between quantum states). In these respects, graphene has potential advantages⁵ over other materials that are being investigated as hosts for spin qubits, such as gallium arsenide or silicon: electric currents flowing through a graphene sheet provide a means of coupling and manipulating spins; and the two main sources of decoherence are minimal in graphene. These sources of decoherence are the coupling between an electron's spin and its orbital motion (which is weak in graphene), and interactions of electron spins with atoms that have nuclear spins (the concentration of which is low in graphene).

Why has it been so difficult to observe magnetic edge states experimentally? The electronic and magnetic properties of graphene nanoribbons correlate closely with the structures of their edges, and are sensitive to even minute numbers of defects. Isolating a sufficient number of nanoribbons that have perfect zigzag edges to enable their magnetic characterization is extremely challenging, and so the data from such studies⁶ are scarce and inconclusive. Experiments performed on single graphene layers prepared *in situ* under a high vacuum have revealed the formation of local electronic states at edges, but did not provide any evidence of magnetism⁷.

By expanding a previously developed



50 Years Ago

One of the best ways to spread plant diseases is through the sale and shipment of seed. In some cases, such as celery leaf rust, only one infected plant in 10,000 is needed to cause an epidemic in the crop. Particularly critical are fungal diseases lodged within the seed ... The problem in dealing with these diseases has been to kill the fungus but not the seed. This type of disease can now be completely eliminated by a process developed at the National Vegetable Research Station ... The treatment is first to soak seed for twenty-four hours in a solution containing 0.2 per cent of the fungicide 'Thiram' at 30 °C. The seed is then dried by driving air through it for several hours. So far this treatment has been found to give complete control in eleven commercially important plant species with infections involving eighteen different seed-borne diseases.

From *Nature* 1 June 1968

100 Years Ago

The trustees of the British Museum have published a report on an investigation carried out ... to ascertain how and when the infestation of Army biscuits by flour-moths takes place, and whether any steps can be taken to prevent this. A list is given of eight species of beetles and four Pyralid moths that were actually found in the tins of biscuits examined. But by far the most serious pest was the moth *Ephestia kühniella* ... Evidence is adduced indicating that Central America is probably the original home of *E. kühniella*, the so-called Mediterranean flour-moth. The examination of various intact airtight tins showed that the biscuits contained in them were infested, thus indicating that the moths had gained access to them in the factory prior to packing.

From *Nature* 30 May 1918

chemical method⁷, Slota *et al.* synthesized graphene nanoribbons in solution that have uniform widths and zigzag edges. The authors attached nitronyl nitroxide molecules — chemically robust organic radicals, which are magnetic because they carry an unpaired electron — to specific edge sites (Fig. 1). This method produces large amounts (milligram quantities) of chemically stable graphene nanoribbons that can be studied using conventional spectroscopic techniques. The authors show that the electron spins at the radicals induce a spin density at the edge carbon sites where the radicals are bonded, and therefore induce magnetic edge states. This trick is akin to moving a row of corks on a string up and down on the surface of a pool to induce ordered water oscillations at the pool's edge; not only do the corks induce waves, but they also make them easier to visualize.

Besides proving the existence of magnetic edge states, Slota and colleagues' experiments provide the first direct determination of the strength of the tiny spin–orbit coupling in their system. These findings will help to validate theoretical models of the electronic structure of graphene and its edge states⁸.

The authors also measured the characteristic rates at which spins relax (reach equilibrium with the graphene lattice) and the time taken for them to lose coherence. The measured decoherence times are roughly one microsecond at room temperature — which is promising, because it means that spin coherence is preserved for much longer than has previously been measured in graphene electronic devices. A plausible explanation for this is that the graphene nanoribbons are free from the structural randomness and extrinsic effects (such as spin scattering caused by connecting graphene to electrodes) that have suppressed spin coherence in other systems⁹. Slota *et al.* find that decoherence in their nanoribbons seems to be mainly associated with interactions of the electron spins with nuclear spins in the radical molecule. This is good news, because chemical methods are available to reduce the concentration of nuclear spins, or to make spin qubits insensitive to the magnetic noise generated by nuclear spins¹⁰.

Finally, the authors showed that unpaired electrons at the radicals interact with the edge spins. These interactions might allow graphene to be used as a coherent communication channel between different radical spins, and might therefore serve as the basis of the two-qubit logic gates necessary for a quantum computer.

Slota *et al.* show that the attachment of magnetic molecules to graphene creates coherent magnetic states on it, nicely complementing previously reported experiments¹¹ that showed how graphene influences the electron spins on molecules deposited on it. However, in the authors' system, electron spin is 'injected' into the nanoribbons from the radical molecules — so the intrinsic magnetism of graphene edges remains to be investigated.

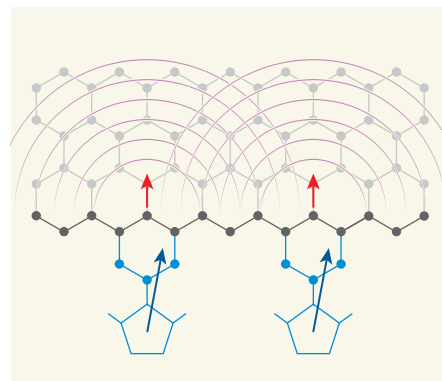


Figure 1 | Electron spins at graphene edges. Slota *et al.*² have made ribbons of graphene (grey) that have zigzag edges (black), and free-radical molecules (blue) attached at specific sites. Each molecule has an unpaired electron, which has an associated quantum property known as spin (blue arrows). The molecules stabilize the carbon nanoribbons and perturb electrons in the graphene at the edges (purple ripples), generating electron spins at the edges (red arrows).

One way to explore this would be to attach non-magnetic molecules, rather than free radicals, to the graphene edges.

A formidable challenge in the development of the reported nanoribbons for quantum computers will be to design a system that can manipulate and read out each qubit in a nanoribbon, and that can switch interactions between qubits on and off, in a way that also allows the computer to expand to incorporate more qubits without losing control of them. This will probably require graphene nanosheets to be coupled to a solid-state device, so it remains to be seen how the effects of coupling to the device will affect spin coherence.

Moreover, if the strength of the spin–orbit coupling of edge-modified graphene nanoribbons can be increased, then the spin at the attached molecules could be manipulated using an electric field. Such strengthening might be achieved by replacing the organic radicals with molecular metal complexes — which would require new chemical methods. It therefore seems that chemists hold the key to technologies and scientific discoveries involving magnetic graphene. ■

Fernando Luis is at the Instituto de Ciencia de Materiales de Aragón, CSIC – University of Zaragoza, Zaragoza 50009, Spain.

Eugenio Coronado is at the Instituto de Ciencia Molecular, University of Valencia, Paterna 46980, Spain.

e-mails: fluis@unizar.es;

eugenio.coronado@uv.es

1. Son, Y.-W., Cohen, M. L. & Louie, S. G. *Nature* **444**, 347–349 (2006).
2. Slota, M. *et al.* *Nature* **557**, 691–695 (2018).
3. Han, W., Kawakami, R. K., Gmitra, M. & Fabian, J. *Nature Nanotechnol.* **9**, 794–807 (2014).
4. Pesin, D. & MacDonald, A. H. *Nature Mater.* **11**, 409–416 (2012).
5. Recher, P. & Trauzettel, B. *Nanotechnology* **21**,

- 302001 (2010).
 6. Rao, S. S. *et al.* *ACS Nano* **6**, 7615–7623 (2012).
 7. Ruffieux, P. *et al.* *Nature* **531**, 489–492 (2016).
 8. Min, H. *et al.* *Phys. Rev. B* **74**, 165310 (2006).
 9. Drögerler, M. *et al.* *Nano Lett.* **16**, 3533–3539 (2016).
 10. Shiddiq, M. *et al.* *Nature* **531**, 348–351 (2016).
 11. Cervetti, C. *et al.* *Nature Mater.* **15**, 164–168 (2016).

SYNTHETIC BIOLOGY

Yeast shuffles towards a diverse future

A redesigned yeast genome is being constructed to allow it to be extensively rearranged on demand. A suite of studies reveals the versatility of the genome-shuffling system, and shows how it could be used for biotechnology applications.

JEE LOON FOO & MATTHEW WOOK CHANG

A global consortium of scientists is well on the way to making a synthetic genome for the yeast *Saccharomyces cerevisiae*¹ — the first synthetic genome for a member of the group of organisms known as eukaryotes, which includes plants, animals and fungi. Embedded within the extensively redesigned ‘version 2.0’ genome of *S. cerevisiae* (Sc2.0) are DNA sequences that form part of a system known as Synthetic Chromosome Rearrangement and Modification by LoxP-mediated Evolution (SCRaMbLE). This system allows extensive reorganization of the genome to be triggered on demand, generating Sc2.0 variants that have diverse genetic make-ups and characteristics. Sc2.0 is therefore a versatile platform that can be easily modified and evolved to produce yeasts that have desired attributes². A collection of seven papers^{3–9} published in *Nature Communications* demonstrates the immense potential of Sc2.0 for engineering and understanding yeast.

To enable SCRaMbLE, a palindromic DNA sequence known as *loxPsym* is inserted after

every non-essential gene in the synthetic genome. In the presence of the enzyme Cre recombinase, the *loxPsym* sites undergo recombination with each other — that is, the *loxPsym* sequences break in the middle, and the broken ends can then join up with any other available *loxPsym* ends. This process results in genes being randomly deleted, inverted, relocated and duplicated.

In the original design of the SCRaMbLE system¹⁰, Cre recombinase was produced only once during the lifetime of a cell, and was fused to a protein domain that binds oestradiol molecules — which allowed the enzyme to be activated by adding oestradiol to the yeast’s growth medium, providing an on–off switch for genome rearrangement (Fig. 1). However, some ‘background’ genome rearrangement occurred even without oestradiol activation. This version of SCRaMbLE was functional^{11,12}, but four of the new papers now report improvements to the system.

Shen *et al.*³ have modified SCRaMbLE to produce multiple pulses of Cre recombinase (instead of just one per lifetime) to increase rearrangement events while reducing

background Cre recombinase activity. Jia *et al.*⁴ have developed a SCRaMbLE variant in which both oestradiol and galactose molecules are required to activate rearrangement, also reducing background rearrangement. Hochrein *et al.*⁵ have engineered Cre recombinase so that it is activated by red light, providing a new way to control SCRaMbLE. And Luo *et al.*⁶ have introduced a reporter DNA sequence into a synthetic yeast strain, which allows cells that have undergone SCRaMbLE-induced genome rearrangement to be easily distinguished from those that have not. All four improvements facilitate effective and efficient implementation of SCRaMbLE.

An important application of SCRaMbLE is to generate genetically diverse pools of yeast mutants from which strains that have industrially valuable characteristics can be isolated. For example, yeasts can be genetically engineered to produce useful compounds, and Blount *et al.*⁷ show that SCRaMbLE can generate yeast strains that produce antibiotics (violacein or penicillin) in greater quantities than could be achieved without SCRaMbLE. Blount and colleagues also used the system to produce yeast strains that use the sugar xylose for growth more effectively than strains produced without SCRaMbLE; xylose is poorly used by wild-type yeast, but is abundant in biomass and is therefore an attractive alternative to the sugars normally used to feed yeast in industrial applications. And Luo *et al.* have used their SCRaMbLE variant to accelerate the isolation of yeast strains that are tolerant to various stress factors, such as ethanol, heat and acetic acid.

Jia and co-workers report that production of β -carotene molecules can be drastically increased if SCRaMbLE is used in diploid yeasts, which have two copies of the genome, instead of haploids, which have a single copy. Similarly, Shen *et al.* used SCRaMbLE in diploids to improve the heat or caffeine tolerance of hybrid yeasts (organisms produced by crossing two different yeast species or subspecies).

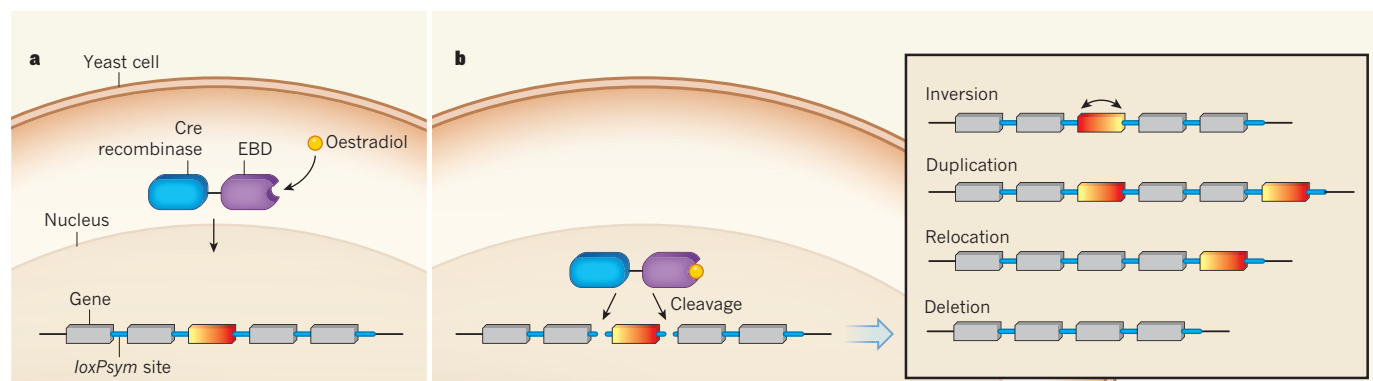


Figure 1 | Genome rearrangement on demand. A synthetic genome of the yeast *Saccharomyces cerevisiae* is being constructed that allows the genome to be rearranged using a system known as Synthetic Chromosome Rearrangement and Modification by LoxP-mediated Evolution (SCRaMbLE). In the first version of this system, a palindromic DNA sequence known as *loxPsym* is inserted after every non-essential gene, and a protein consisting of the enzyme Cre recombinase attached to an oestradiol-binding domain (EBD)

resides in the yeast cytoplasm. When the protein is activated by the binding of an oestradiol molecule, it moves into the nucleus (a) where it cleaves the *loxPsym* sequences (b). The broken ends of *loxPsym* can then join up with any other available *loxPsym* ends, rearranging the genome. This process results in genes (such as the coloured rectangle) being randomly inverted, duplicated, relocated or deleted. Seven papers^{3–9} now report improvements and applications of the SCRaMbLE system.

Both groups observed genome rearrangements in diploids that involved the deletion of one copy of essential genes. The presence of such rearrangements in improved diploid strains shows that, compared to haploids, diploids are more robust to deleterious deletions during SCRaMbLE. This in turn allows a greater number of beneficial rearrangements to be manifested. Although it is premature to claim that SCRaMbLE is a universal tool for engineering yeast, taken together, the various findings^{3–7} certainly show that it has great potential for generating yeasts for a wide range of purposes.

Wu *et al.*⁸ have taken SCRaMbLE out of cells and used it *in vitro* with purified Cre recombinase to generate different genetic arrangements of the β -carotene biosynthetic pathway. They thus discovered arrangements that increase β -carotene production compared with the original pathway. By contrast, Liu *et al.*⁹ used an *in vitro* method involving recombinase enzymes separate from the SCRaMbLE system, to rapidly generate different versions of β -carotene- and violacein-producing pathways and to identify highly productive ones. They then flanked the DNA sequences of the best pathways with *loxP* sites, and used SCRaMbLE to randomly incorporate the pathways at *loxP* sites in the synthetic yeast genome. SCRaMbLE concurrently rearranged the resulting genomes, allowing yeast strains to be optimized for the production of the desired compounds. These two papers illustrate the versatility of the basic SCRaMbLE concept and how it can be used in innovative ways.

So where next for Sc 2.0? So far, six synthetic chromosomes of Sc 2.0 have been completed¹³, and consortium members are working full-time to construct the remaining ten. The seven new papers show that researchers are eager to work with the newly available synthetic chromosomes to see how SCRaMbLE techniques can generate useful yeast variants and improve our understanding of the fundamental processes and properties of yeast. Thousands of *loxP* sites will be present in the fully assembled Sc 2.0 genome, and so the number of genomic structures that can be generated by SCRaMbLE is immense — which suggests that it should be possible to produce a yeast variant that displays any desired set of characteristics.

Nevertheless, SCRaMbLE systems are still in their infancy. Further improvements are needed, along with tools that maximize the potential of SCRaMbLE-based techniques. For example, the screening of SCRaMbLE-modified yeast has generally relied on visible cues, such as growth rate and colour (both β -carotene and violacein are pigments that colour the yeast cells). Luo and colleagues' reporter offers a useful new screening tool, but high-throughput methods are also needed that can identify yeast strains that produce large amounts of colourless chemicals. Crucially, the characterization of genetic rearrangements relies heavily on whole-genome sequencing. The development of

more-efficient, cheaper sequencing techniques would allow more strains to be sequenced than is currently possible, to work out and study changes in the genome. Given the promising early results and synergy among the members of the Sc2.0 consortium, the establishment of SCRaMbLE as a staple tool for engineering yeast is highly anticipated. ■

Jeon Loon Foo and Matthew Wook Chang are in the Department of Biochemistry, Yong Loo Lin School of Medicine, and in the NUS Synthetic Biology for Clinical and Technological Innovation (SynCTI), Centre for Life Sciences, National University of Singapore, Singapore 117456.
e-mail: bchcmw@nus.edu.sg

1. Burgess, D. J. *Nature. Rev. Genet.* **18**, 274 (2017).

In retrospect

A catalyst for 50 years of cancer research

In 1968, a defect in DNA repair was found to underlie a disorder that makes people extremely sensitive to sunlight. This finding continues to influence research into the origins, diagnosis and treatment of cancer.

RICHARD D. WOOD

Some people are born with exceptional sensitivity to sunlight. Fifty years ago, writing in *Nature*, the biologist James Cleaver¹ reported a study of one such condition, and concluded that a failure of DNA repair was related to the extreme susceptibility of affected individuals to skin cancer. This was the first description of defective DNA repair in a genetically inherited disorder that makes people prone to cancer. The concepts that developed from this work now permeate research into the genetic origins of cancer and its treatment.

Starting in the 1870s, the Viennese dermatologist Moritz Kaposi performed pioneering work that defined a rare disorder characterized by high sensitivity to sunlight. Young patients were severely burned by brief exposure to the sun and acquired frequent skin lesions, and some had a high incidence of skin tumours. Kaposi dubbed the condition xeroderma pigmentosum² (XP), using the Greek words for dry, pigmented skin — one of the symptoms of the disease. He recognized that this was a hereditary syndrome, but the underlying cause was not obvious.

Little research into XP was then done until the 1960s, when a process called nucleotide excision repair was discovered in bacteria^{3–5}. In this process, enzymes clip out segments of

- Enyeart, P. J. & Ellington, A. D. *Nature* **477**, 413–414 (2011).
- Shen, M. J. *et al. Nature Commun.* <https://doi.org/10.1038/s41467-018-04157-0> (2018).
- Jia, B. *et al. Nature Commun.* <https://doi.org/10.1038/s41467-018-03084-4> (2018).
- Hochrein, L., Mitchell, L. A., Schulz, K., Messerschmidt, K. & Mueller-Roeber, B. *Nature Commun.* <https://doi.org/10.1038/s41467-017-02208-6> (2018).
- Luo, Z. *et al. Nature Commun.* <https://doi.org/10.1038/s41467-017-00806-y> (2018).
- Blount, B. A. *et al. Nature Commun.* <https://doi.org/10.1038/s41467-018-03143-w> (2018).
- Wu, Y. *et al. Nature Commun.* <https://doi.org/10.1038/s41467-018-03743-6> (2018).
- Liu, W. *et al. Nature Commun.* <https://doi.org/10.1038/s41467-018-04254-0> (2018).
- Dymond, J. S. *et al. Nature* **477**, 471–476 (2011).
- Annaluru, N. *et al. Science* **344**, 55–58 (2014).
- Shen, Y. *et al. Genome Res.* **26**, 36–49 (2016).
- van der Sloot, A. & Tyers, M. *Mol. Cell* **66**, 441–443 (2017).

This article was published online on 22 May 2018.

DNA that have been damaged by light and replace them with fresh, undamaged DNA. Mutant bacterial strains were isolated that could be killed by low doses of ultraviolet radiation, and some of these were found to be unable to carry out excision repair^{4,5}.

These concepts of DNA repair were then extended to human cells. By 1964, the biologists Robert Painter and Ronald Rasmussen had discovered that UV irradiation of mammalian cells led to a phenomenon that they interpreted as excision repair⁶. In their experiments, cultured human cells were supplied with radioactive molecules (bases) that could be incorporated into DNA. The cells were observed to incorporate new bases after UV irradiation, even when they were not duplicating their genomes, indicating that UV-damaged DNA was being replaced.

In 1967, Cleaver joined Painter's laboratory in San Francisco as a postdoctoral fellow. Cleaver had obtained his PhD at the University of Cambridge, UK, where he had been using radioactive bases to label DNA in human cells. In April of that year, Cleaver read a newspaper article in the *San Francisco Chronicle* that mentioned research showing that skin cells grown from patients with XP were extraordinarily sensitive to UV radiation⁷. Cleaver raised with Painter the idea that XP might involve a mutation that causes DNA repair to be defective, and suggested investigating this

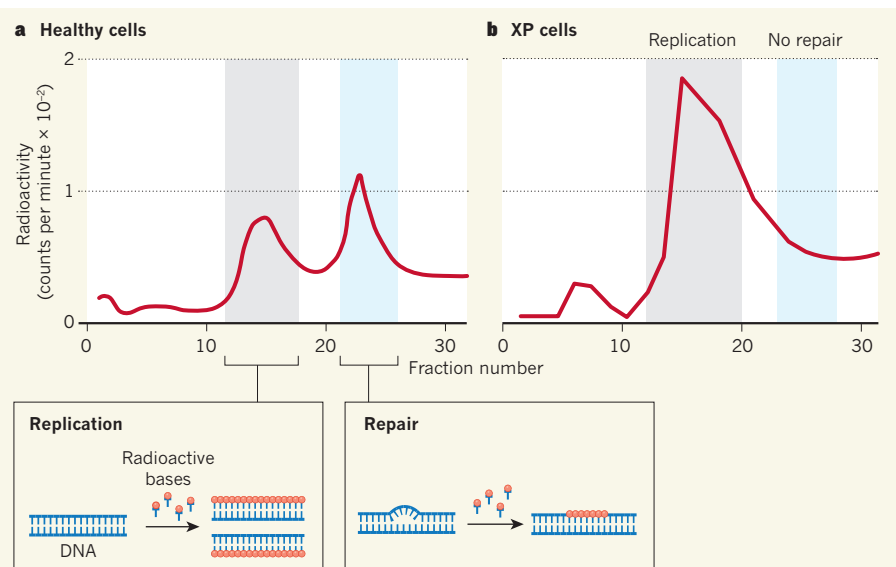


Figure 1 | Evidence of defective DNA repair in cells from people with xeroderma pigmentosum. People born with the condition known as xeroderma pigmentosum (XP) are extremely sensitive to sunlight and are prone to skin cancer. In 1968, Cleaver¹ reported experiments in which cultured cells from a healthy individual and from people with XP were irradiated with ultraviolet light to cause DNA damage and then analysed to see whether the cells incorporated radioactive molecules (bases) into their DNA. **a**, For the healthy cells, plots of measured radioactivity for different fractions of DNA revealed distinct peaks associated with DNA replication and DNA repair. **b**, By contrast, the XP cells lacked the repair-associated peak. This was the first evidence that defective DNA repair underpins a genetically inherited disorder that makes people susceptible to cancer.

clear that many of the XP-associated genes have functions in addition to excision repair, and several are essential for life^{13,14}. This means that only mild disablement of the functions of some XP genes can be tolerated.

Although XP is a rare disease (fewer than 1 person in 250,000 is affected in the United States and Western Europe)¹³, the consequences of mutations in XP genes are being explored widely. For example, a recent analysis found that mutations in the *XPB* gene (also known as *ERCC2*) are fairly frequent in cancer and might modulate individual responses to treatment¹⁷. There is also active research aimed at suppressing the action of XP proteins in tumour cells, to improve the effectiveness of chemotherapies that damage DNA¹⁸.

There is still no cure for XP, but intensive research into the disease means that an early diagnosis can be made. People with XP can then be protected rigorously from sunlight, allowing them a greater quality of life and longer life expectancy than was previously possible. XP societies in the United States and Europe provide support for affected children, with retreats such as Camp Sundown and Owl Patrol. Retinoid compounds can reduce the incidence of skin tumours¹⁴, and dietary interventions might improve the prospects for people with XP and related disorders¹⁹. More broadly, Cleaver's discovery of the DNA-repair defect in XP continues to spawn vigorous research into responses to environmental DNA damage that applies not only to humans, but to every organism on the planet. ■

Richard D. Wood is in the Department of Epigenetics and Molecular Carcinogenesis, The University of Texas MD Anderson Cancer Center, Smithville, Texas 78957, USA. e-mail: rwood@mdanderson.org

1. Cleaver, J. E. *Nature* **218**, 652–656 (1968).
2. Kaposi, M. *Wien. Med. Jb.* (October 1882).
3. Pettijohn, D. & Hanawalt, P. J. *Mol. Biol.* **9**, 395–410 (1964).
4. Boyce, R. P. & Howard-Flanders, P. *Proc. Natl Acad. Sci. USA* **51**, 293–300 (1964).
5. Setlow, R. B. & Carrier, W. L. *Proc. Natl Acad. Sci. USA* **51**, 226–231 (1964).
6. Rasmussen, R. E. & Painter, R. B. *Nature* **203**, 1360–1362 (1964).
7. Perlman, D. *San Francisco Chronicle* 12 April p.4 (1967).
8. Cleaver, J. E. *DNA Repair* **10**, 906–914 (2011).
9. Wood, R. D. *Annu. Rev. Biochem.* **65**, 135–167 (1996).
10. Sancar, A. *Angew. Chem. Int. Edn* **55**, 8502–8527 (2016).
11. Lederberg, J. *The Washington Post and Times-Herald* 8 June p.A13 (1968).
12. de Weerd-Kastelein, E. A., Keijzer, W. & Bootsma, D. *Nature N. Biol.* **238**, 80–83 (1972).
13. Lehmann, A. R., McGibbon, D. & Stefanini, M. *Orphanet J. Rare Dis.* **6**, 70 (2011).
14. DiGiovanna, J. J. & Kraemer, K. H. *J. Invest. Dermatol.* **132**, 785–796 (2012).
15. Aboussekhr, A. et al. *Cell* **80**, 859–868 (1995).
16. Mu, D., Hsu, D. S. & Sancar, A. *J. Biol. Chem.* **271**, 8285–8294 (1996).
17. Knijnenburg, T. A. et al. *Cell Rep.* **23**, 239–254 (2018).
18. Gavande, N. S. et al. *Pharmacol. Ther.* **160**, 65–83 (2016).
19. Vermeij, W. P. et al. *Nature* **537**, 427–431 (2016).

possibility. Painter replied: “It’s a crazy idea, but at your stage what have you got to lose!”⁸

Cleaver acquired cultures of growing skin cells from people with XP, and applied newly developed techniques^{3,6} to determine whether the cells were capable of excision repair. The results clearly showed that DNA repair was defective in XP cells that had been damaged by UV irradiation (Fig. 1). Painter was a generous mentor, and encouraged his junior colleague to pursue this major discovery independently. Cleaver's results were published in *Nature* on 18 May 1968.

The paper's conclusions were strong. Cleaver used two completely different methods to show that DNA repair in XP cells is defective, using cells from three patients clinically verified to have XP, and control cells taken from a patient with an unrelated hereditary disorder and from a healthy individual. The results suggested that XP is not a homogeneous disease, because cell lines from different individuals exhibited different levels of DNA repair. There was no indication, however, of which step was affected in the repair process, or which genes might be altered. Cleaver estimated that about 70 DNA bases were incorporated in each repair event — not far from the actual number of about 30 bases per repair event obtained later using more-precise methods^{9,10}.

The publication generated immediate excitement. DNA repair had previously been considered a somewhat obscure topic, but Cleaver showed that it had a key role in human health. The Nobel-prizewinning molecular biologist Joshua Lederberg penned an editorial

in *The Washington Post* highlighting this important example of fundamental research that turned out to be relevant to disease¹¹. J. Michael Bishop, who won a Nobel prize in 1989 for his work on oncogenes, which have the potential to cause cancer, was also influenced by the finding. He wrote⁸: “While I was still in medical school, James Cleaver recognized xeroderma pigmentosum as a deficiency in the repair of DNA damage caused by ultraviolet light... I have been a believer in the somatic mutation hypothesis of cancer ever since”. Somatic mutations are caused by DNA damage and copying errors in the genes of tumour cells as cancer progresses. Cleaver's paper helped to stimulate the worldwide explosion of DNA-repair research that started in the 1970s⁸.

Cleaver's results were soon confirmed and extended by laboratories around the world. In 1972, it was reported that XP is a genetically complex disease¹², and it is now known that alterations in eight different genes can give rise to it^{13,14}. Seven of these genes encode components of the molecular machinery that performs excision repair; this machinery was biochemically reconstituted *in vitro* in the 1990s^{15,16}. One form of XP, however, is caused by abnormal DNA synthesis after UV irradiation¹³, rather than by a problem in excision repair.

Specific defects in DNA repair are now known to be associated with major neurological and developmental abnormalities in other UV-sensitivity disorders, including Cockayne syndrome^{13,14}. More broadly, it has become

Emerging trends in global freshwater availability

M. Rodell^{1*}, J. S. Famiglietti^{2,5}, D. N. Wiese², J. T. Reager², H. K. Beaudoin^{1,3}, F. W. Landerer² & M.-H. Lo⁴

Freshwater availability is changing worldwide. Here we quantify 34 trends in terrestrial water storage observed by the Gravity Recovery and Climate Experiment (GRACE) satellites during 2002–2016 and categorize their drivers as natural interannual variability, unsustainable groundwater consumption, climate change or combinations thereof. Several of these trends had been lacking thorough investigation and attribution, including massive changes in northwestern China and the Okavango Delta. Others are consistent with climate model predictions. This observation-based assessment of how the world's water landscape is responding to human impacts and climate variations provides a blueprint for evaluating and predicting emerging threats to water and food security.

Groundwater, soil moisture, surface waters, snow and ice are dynamic components of the terrestrial water cycle^{1–3}. Although they are not static on an annual basis (as early water-budget analyses supposed), in the absence of hydroclimatic shifts or substantial anthropogenic stresses they typically remain range-bound. Recent studies have identified locations where terrestrial water storage (TWS; the sum of these five components) appears to be trending below previous ranges, notably where ice sheets or glaciers are diminishing in response to climate change^{4,5} and where groundwater is being withdrawn at an unsustainable rate^{6–8}.

Accurate accounting of changes in freshwater availability is essential for predicting regional food supplies, human and ecosystem health, energy generation and social unrest. Groundwater is particularly difficult to monitor and manage because aquifers are vast and unseen, yet groundwater meets the domestic needs of roughly half of the world's population⁹ and boosts food supply by providing for 38% of global consumptive irrigation water demand¹⁰. Nearly two-thirds of terrestrial aquatic habitats are being increasingly threatened¹¹, while the precipitation and river discharge that support them are becoming more variable¹². A recent study¹¹ estimates that almost 5 billion people live in areas where threats to water security are likely—a situation that will only be exacerbated by climate change, population growth and human activities. Therefore, the key environmental challenge of the 21st century may be the globally sustainable management of water resources.

Much of our knowledge of past and current freshwater availability comes from a limited set of ground-based, point observations. Assessing changes in hydrologic conditions at the global scale is exceedingly difficult using in situ measurements alone, owing to the cost of installing and maintaining instrument networks, the presence of gaps in those networks and the lack of digitization and sharing of existing data¹³. Satellite remote sensing has proven crucial to monitoring water storage and fluxes in a changing world, enabling a truly global perspective that spans political boundaries¹⁴. In particular, since its launch in 2002, the GRACE mission¹⁵ has tracked ice-sheet and glacier ablation, groundwater depletion and other TWS changes^{16–19}. On a monthly basis GRACE can resolve TWS changes with sufficient accuracy over scales that range from approximately 200,000 km² at low latitudes to about 90,000 km² near the poles¹. However, owing to GRACE's coarse spatial resolution, the inability to partition component mass changes and the brevity of the time series, proper

attribution of the TWS changes requires comprehensive examination of all available auxiliary information and data, which has never before been performed at the global scale.

Here we map TWS change rates around the globe based on 14 years (April 2002 – March 2016) of GRACE observations (Fig. 1). The GRACE data were processed using an advanced mass concentration²⁰ ('mascon') approach that enables improved signal resolution relative to the standard spherical-harmonic technique²¹. Best-fit linear rates of change after removing the seasonal cycle (referred to herein as 'apparent trends') are presented in Table 1 for 34 study regions. For context, the largest man-made reservoir in the USA, Lake Mead, has a capacity of about 32 Gt; during the study period, all but one of the 34 regions lost or gained more water than that, and eleven of them lost or gained more than ten times that amount. The reported uncertainty bounds are typically low because the process of removing glacial isostatic adjustment (GIA) signals is the only major source of error in the secular signal. Therefore, low uncertainty does not, on its own, imply that the apparent trends existed before the GRACE period or will continue into the future. The coefficient of determination (r^2), which represents the 'goodness of fit' of the regressed linear trends, is included in Table 1 to quantify the strength of the apparent trends relative to non-secular interannual variability. It is hence a useful, but by no means conclusive, piece of evidence that can be used to predict whether the trend will be fleeting or enduring, reflecting the cohesiveness of the TWS time series tendencies, as shown in Extended Data Fig. 1–4. We attribute the trends to natural variability, direct human impacts or climate change and forecast the likelihood that they will continue on the basis of 1979–2016 precipitation data from the Global Precipitation Climatology Project version 2.3 (GPCP)²² (see Extended Data Figs. 5–8), an irrigated area map²³, satellite-based lake-level altimetry time series²⁴, Landsat imagery and published reports of human activities including agriculture, mining, reservoir operations and inter-basin water transfers. Further, for each region we provide the median climate model prediction of precipitation changes between 1986–2005 and 2081–2100 using the Representative Concentration Pathways 8.5 W m⁻² (RCP8.5; 8.5 W m⁻² radiative forcing in 2100 relative to pre-industrial levels) greenhouse gas emissions scenario from the Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report²⁵. We chose the high-end ('business as usual')

¹Hydrological Sciences Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD, USA. ²Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA. ³Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD, USA. ⁴Department of Atmospheric Sciences, National Taiwan University, Taipei, Taiwan. ⁵Present address: Global Institute for Water Security, School of Environment and Sustainability, and Department of Geography and Planning, University of Saskatchewan, Saskatoon, Canada. *e-mail: Matthew.Rodell@nasa.gov

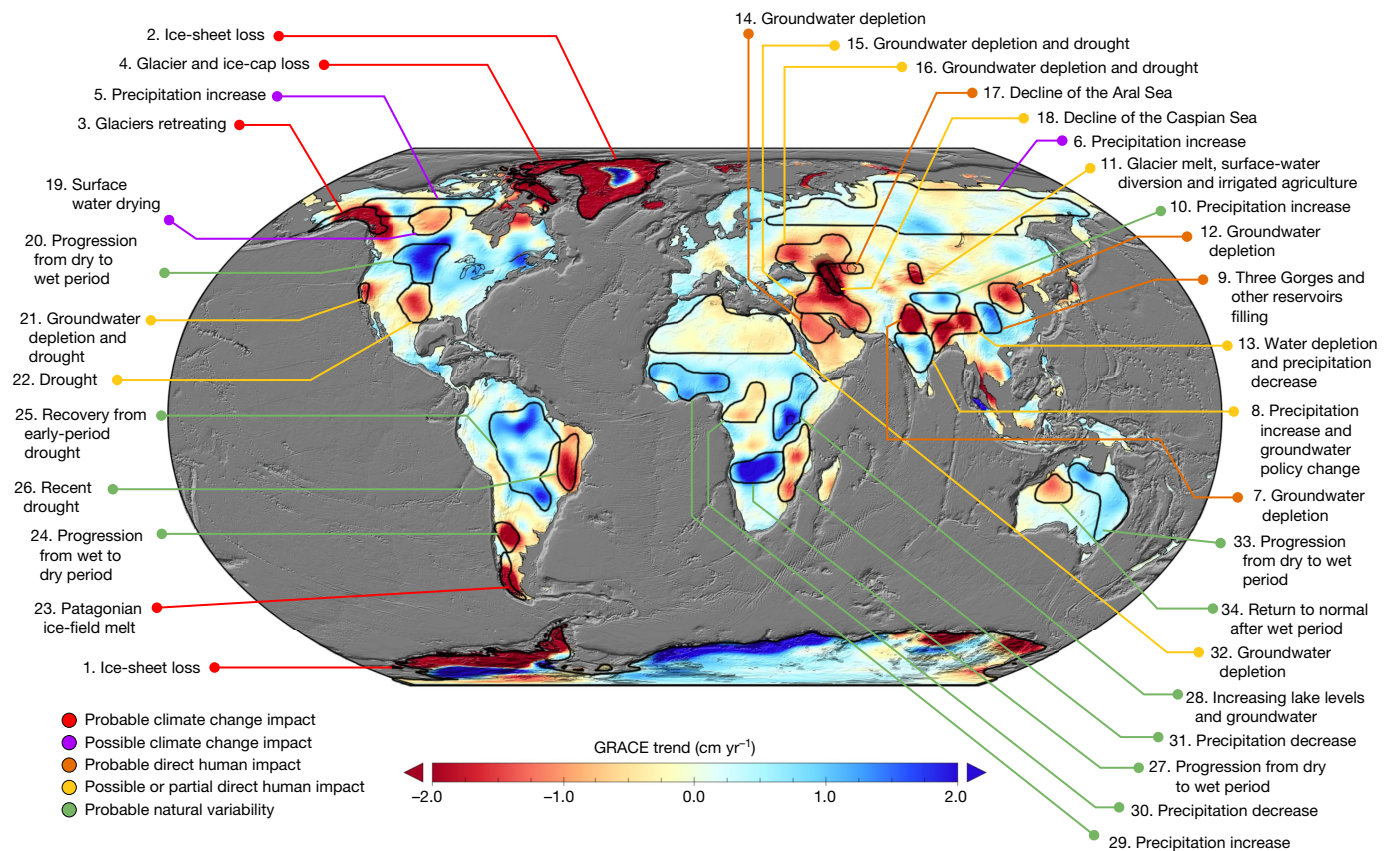


Fig. 1 | Annotated map of TWS trends. Trends in TWS (in centimetres per year) obtained on the basis of GRACE observations from April 2002 to March 2016. The cause of the trend in each outlined study region is briefly explained and colour-coded by category. The trend map was smoothed

with a 150-km-radius Gaussian filter for the purpose of visualization; however, all calculations were performed at the native 3° resolution of the data product.

scenario because it accentuates regional differences, which are more important for this analysis than absolute magnitudes are. Figure 2 presents maps of the IPCC, GPCP and irrigated-area data.

Global scale

By far the largest TWS trends occur in Antarctica (region 1; $-127.6 \pm 39.9 \text{ Gt yr}^{-1}$ averaged over the continent), Greenland (region 2; $-279.0 \pm 23.2 \text{ Gt yr}^{-1}$), the Gulf of Alaska coast (region 3; $-62.6 \pm 8.2 \text{ Gt yr}^{-1}$) and the Canadian archipelago (region 4; $-74.6 \pm 4.1 \text{ Gt yr}^{-1}$), where the warming climate continues to drive rapid ice-sheet and glacier ablation^{4,5,26,27}. Positive trends in sub-regions of Antarctica and Greenland result from increasing snow accumulation²⁸ and millennial-scale dynamic thickening processes^{29,30}. Excluding those four ice-covered regions, one of the most striking aspects of changing TWS illuminated by Fig. 1 is that freshwater seems to be accumulating in far-northern North America (region 5) and Eurasia (region 6) and in the wet tropics, whereas the greatest non-frozen-freshwater losses have occurred at mid-latitudes^{8,31}. The observed trends are consistent with increasing rates of northern high-latitude precipitation during the study period and with the prediction of IPCC models that precipitation generally will decrease in mid-latitudes and increase in low and high latitudes by the end of this century²⁵. They also complement recent studies that identify increasing rates of precipitation in the tropics and increasing water storage and river discharge in the high Arctic^{12,32}. However, because the rates of TWS change ($0.45 \pm 0.43 \text{ cm yr}^{-1}$ and $0.17 \pm 0.12 \text{ cm yr}^{-1}$ in regions 5 and 6, respectively) and the coefficients of determination (0.52 and 0.10, correspondingly) are small, while GIA-related errors are relatively large, we cannot state definitively that these high-latitude tendencies are real trends.

A second characteristic of the map is that it reveals a clear 'human fingerprint' on the global water cycle. As seen in Fig. 2, freshwater

is rapidly disappearing in many of the world's irrigated agricultural regions^{6,10,33–38}. A third aspect of the global-trend map is natural inter-annual variability; many of the apparent trends are probably temporary, caused by oscillations between dry and wet periods (themselves driven by El Niño, La Niña and other climatic cycles) during the 14-year study period^{39,40}.

Eurasia

The hotspot in northern India (region 7) was among the first non-polar TWS trends to be revealed by GRACE^{41,42}. It results from groundwater extraction to irrigate crops, including wheat and rice, in a semi-arid climate. Fifty-four per cent of the area is equipped for irrigation. We estimate the rate of TWS depletion to be $19.2 \pm 1.1 \text{ Gt yr}^{-1}$, which is within the range of GRACE-based estimates from previous studies of differently defined northern-India regions^{41–43}. The trend persists despite precipitation being 101% of normal (namely, the 1979–2015 GPCP annual mean for the region) during the study period, with an increasing trend of 15.8 mm yr^{-1} . The fact that extractions already exceed recharge during normal-precipitation years does not bode well for the availability of groundwater during future droughts. The contribution of Himalayan glacier mass loss to the regional trend is minor^{41,42}.

The increasing trend in central and southern India (region 8; $9.4 \pm 0.6 \text{ Gt yr}^{-1}$) probably reflects natural variability of (mostly monsoon) rainfall, which was 104% of normal with an increasing rate of 3.7 mm yr^{-1} (0.4% per year). Although the r^2 value is low (0.24), the TWS and rainfall trends are both consistent with the RCP8.5-predicted 23% precipitation increase by 2100.

The increasing trend in eastern central China (region 9) is caused by a surge in dam construction and subsequent reservoir filling across that region⁴⁴. The best known is the Three Gorges Dam reservoir, which was filled to its design capacity of 39.3 Gt between June 2003 and

Table 1 | TWS trends and supporting information

Region	Location	Area (km ²)	TWS trend (Gt yr ⁻¹)	TWS trend errors (Gt yr ⁻¹)	r ² of TWS trend	Irrigated area (%)	Precipitation trend (mm yr ⁻¹)	Precipitation trend (% yr ⁻¹)	Precipitation percentage of normal (%)	Predicted precipitation change (%)
1	Antarctica	12,397,401	-127.6	39.9	0.93	0	0.82	0.40	96.6	30.9
2	Greenland	2,184,307	-279.0	23.2	0.97	0	8.85	2.09	102.7	39.1
3	Gulf of Alaska coast	716,492	-62.6	8.2	0.93	0	-3.03	-0.25	95.0	21.3
4	Canadian Archipelago	672,413	-74.6	4.1	0.95	0	-5.33	-2.11	94.5	38.3
5	Northern North America	1,350,129	6.1	5.8	0.52	0	2.35	0.73	105.1	26.9
6	Northern Eurasia	8,009,175	13.4	9.7	0.10	0	1.65	0.30	104.4	25.1
7	Northern India	664,169	-19.2	1.1	0.80	54	15.80	2.20	101.0	11.8
8	Central India	1,352,670	9.4	0.6	0.24	51	3.72	0.36	103.7	23.1
9	Eastern Central China	657,375	7.8	1.6	0.78	14	7.33	0.77	99.6	7.9
10	Tibetan Plateau	881,704	7.7	1.4	0.67	0	-1.52	-0.90	104.2	19.7
11	Northwestern China	215,152	-5.5	0.5	0.77	7	1.11	0.57	109.8	15.3
12	North China Plain	876,004	-11.3	1.3	0.63	52	-2.33	-0.37	103.0	19.4
13	Eastern India Region	1,228,839	-23.3	1.9	0.85	25	-9.52	-0.67	96.1	14.7
14	Northwestern Saudi Arabia	841,763	-10.5	1.5	0.92	0	-1.44	-1.31	77.7	-1.4
15	Northern Middle East	2,189,561	-32.1	1.5	0.84	5	-2.80	-0.90	96.3	-8.5
16	Southwestern Russia Region	1,772,712	-18.1	1.3	0.64	15	-5.83	-0.92	96.8	6.2
17	Aral Sea	52,299	-2.2	0.1	0.76	0	2.71	1.17	111.1	5.9
18	Caspian Sea	377,761	-23.7	4.2	0.76	0	-4.37	-1.14	103.4	2.1
19	Central Canada	802,682	-7.0	6.4	0.73	0	0.69	0.17	102.0	16.9
20	Northern Great Plains	1,333,598	20.2	4.8	0.79	3	2.26	0.44	102.0	7.0
21	Southern California	177,996	-4.2	0.4	0.46	18	-8.31	-1.29	89.7	1.2
22	Southern High Plains and eastern Texas	1,105,113	-12.2	3.6	0.44	9	-5.71	-0.76	95.2	-2.8
23	Patagonian ice fields	461,198	-25.7	5.1	0.89	0	-8.01	-0.76	97.1	-6.9
24	Central Argentina ^a	530,661	-8.6	1.2	0.77	4	1.87	0.32	94.2	0.7
25	Central and western Brazil	5,559,805	51.9	9.4	0.39	1	0.61	0.03	100.2	-5.0
26	Eastern Brazil	1,132,450	-16.7	2.9	0.39	1	-16.97	-1.61	97.7	-5.9
27	Okavango Delta	1,589,692	29.5	3.5	0.55	0	-5.21	-0.61	105.3	-8.7
28	Nile headwaters	1,824,276	21.9	3.9	0.56	1	-3.53	-0.30	97.7	11.6
29	Tropical western Africa	2,298,134	24.1	2.1	0.67	1	-0.12	-0.01	103.4	-6.3
30	Northern Congo	1,318,261	-7.2	1.0	0.26	0	-1.55	-0.10	99.1	7.1
31	Southeastern Africa	1,677,719	-12.9	2.3	0.47	0	-3.23	-0.32	95.9	-5.9
32	Northern Africa	6,664,135	-11.7	2.9	0.45	1	-0.12	-0.19	106.7	-12.9
33	Northern & Eastern Australia	2,504,494	19.0	2.8	0.32	3	4.30	0.69	104.6	-6.0
34	Northwestern Australia	1,002,367	-8.9	1.2	0.43	0	-0.39	-0.10	99.1	-0.6

Location; area; GRACE-based TWS trend (April 2002–March 2016) and uncertainty; coefficient of determination (r^2) of the fitted linear trend; percentage of the area equipped for irrigation²³; trend in precipitation²² (January 2002–March 2016) after removing the seasonal cycle; annual mean precipitation (2003–2015) as a fraction of the long-term (1979–2015) annual mean²²; and median precipitation change between the periods 1986–2005 and 2081–2100, predicted using the IPCC high-end greenhouse gas emissions scenario²⁵ for each of the 34 study regions.

^aThe TWS trend in region 24 is for April 2002–February 2010 only.

October 2010⁴⁵. The 14-year regional trend, $7.8 \pm 1.6 \text{ Gt yr}^{-1}$, did not change appreciably after the Three Gorges Dam Reservoir was filled. That can be explained by both the prevalence of other dam projects and the greater precipitation after 2010 (971 mm yr^{-1} ; compared to 928 mm yr^{-1} before 2010). Further, seepage from dams tends to raise the regional water table, which can continue for years before the system equilibrates⁴⁶. If precipitation trends towards an 8% increase by the end of this century, as predicted, then the observed TWS trend may persist even after the current dam building boom, although probably at a slower pace.

Satellite altimetry and Landsat data indicate that the majority of lakes in the Tibetan Plateau have grown in water level and extent during the 2000s owing to a combination of elevated precipitation rates and increased glacier-melt flows⁴⁷, which are difficult to disentangle. From 1997 to 2001 the average annual precipitation in region 10 was 160 mm yr^{-1} , well below the 2002–2015 average of 175 mm yr^{-1} ; thus, the observed increase in TWS ($7.7 \pm 1.4 \text{ Gt yr}^{-1}$) may reflect replenishment after a prolonged dry period. Additional surface-water storage would have been partially offset by glacier retreat and warming-enhanced evaporation. The GIA may further complicate the partitioning of the GRACE-derived mass-change signal over the Tibetan Plateau⁴⁸, but some have argued that the GIA contribution is negligible⁴⁹. The latter study⁴⁹ noted that interannual mass variability in the region during the GRACE period is large relative to the inferred trend⁴⁹. We concur ($r^2 = 0.67$) and conclude that there is no basis for extrapolating the apparent TWS trend into the future. In fact, it appears to have reversed in 2013 (Extended Data Fig. 2). Although RCP8.5

predicts a 20% increase in precipitation by 2100, it is probable that warming-induced glacier-mass losses will begin to exceed surface-water gains, particularly if the fraction of frozen precipitation decreases.

Region 11 lies to the west of the city of Urumqi in northwestern China's Xinjiang province. During the study period, TWS depletion was intense: $-5.5 \pm 0.5 \text{ Gt yr}^{-1}$ from an area of only $215,000 \text{ km}^2$. Precipitation data indicate that drought was a non-factor. The glaciers of the Tien Shan mountain range, whose central third lies within region 11, are melting rapidly⁴⁹, but not rapidly enough to explain all of the mass loss. Groundwater is being withdrawn to support irrigated agriculture across the province^{50,51} and possibly to dewater coal mines⁵². However, region 11 is contained within an endorheic basin. Hence, the additional surface water produced by ice-melt and groundwater abstraction cannot flow far, yet the elevations of the five lakes within that basin either declined or were stable during the study period and GRACE did not detect substantial TWS increases in other parts of the basin. We conclude that region 11 is losing glacier ice and possibly groundwater, which ultimately become evapotranspiration, both in irrigated agricultural areas to the north, south and west of the mountains, as well as through evaporation from the desert floor to the south⁵⁰. Details are provided in Methods.

The vast agricultural region surrounding Beijing (region 12) is heavily irrigated (52%). Previous GRACE-based studies offered a wide range of estimates for groundwater depletion from the North China Plain aquifer (see Methods for details), which is encompassed by region 12 and supports much of that irrigation. Here we estimate a TWS change rate of $-11.3 \pm 1.3 \text{ Gt yr}^{-1}$ for region 12. During the GRACE period,

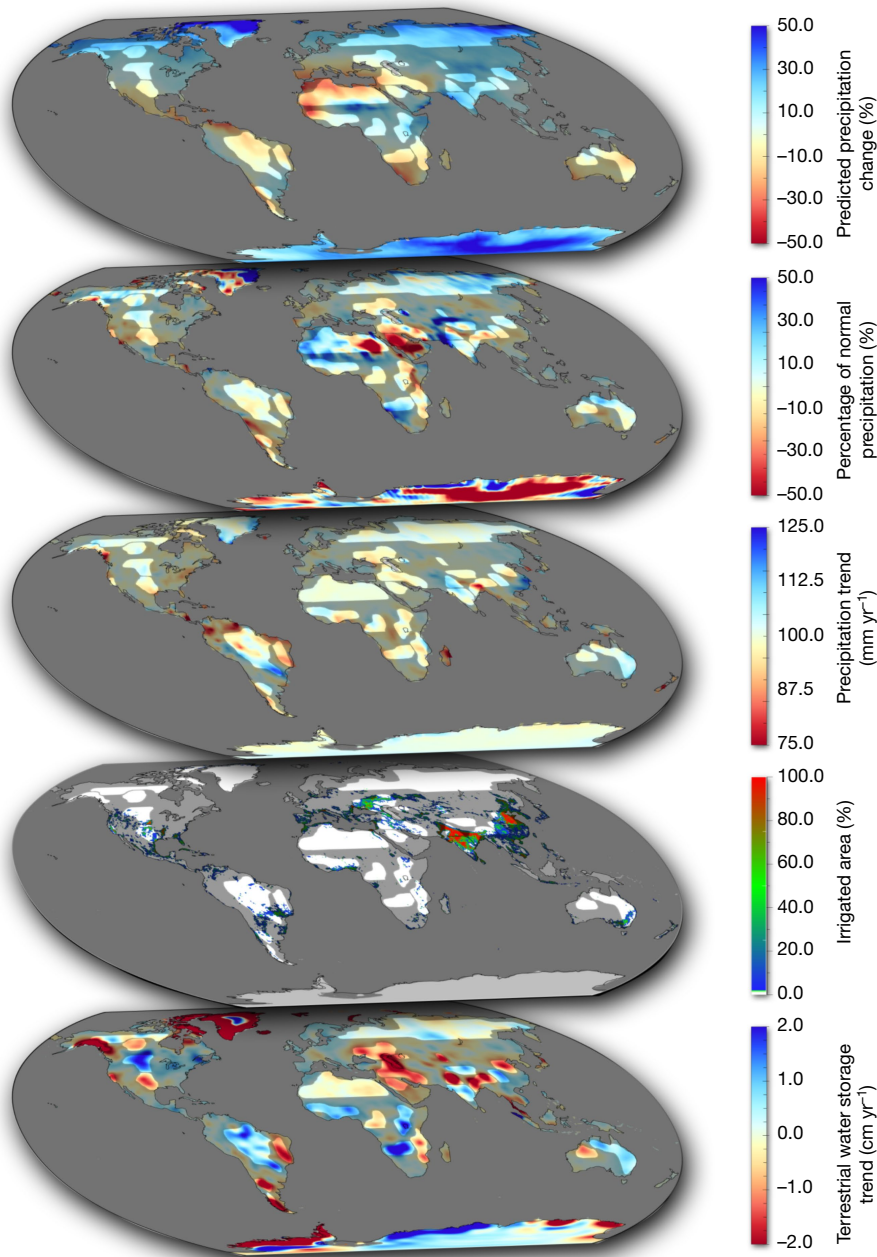


Fig. 2 | Trends in TWS and supporting data maps. (Bottom to top) TWS trends (in centimetres per year); percentage of area equipped for irrigation²³ (%); trend in precipitation²² (in millimetres per year); mean

annual precipitation (2003–2015) as a percentage of the long-term mean²² (in per cent); IPCC-predicted change in precipitation²⁵ (in per cent). Areas outside of the study regions are shaded.

the total annual precipitation was steady at about 10 mm yr^{-1} above the 1979–2015 mean, following two dry years and a wet year during 2001–2003. All evidence suggests that this trend is human-induced and likely to continue until groundwater becomes scarce or regulations are put in place to reduce consumption rates.

The negative trend that extends across East India, Bangladesh, Burma and southern China (region 13), $-23.3 \pm 1.9 \text{ Gt yr}^{-1}$, may be explained by a combination of intense irrigation⁵³ (25%) and a decrease in monsoon season precipitation during the study period. The total annual precipitation was well above normal from 1998 to 2001, resulting in elevated TWS. During the GRACE period, precipitation declined at a rate of -10 mm yr^{-1} (-0.7% per year), and the annual accumulations were below average from 2009 to 2015. This is the third most heavily irrigated of the study regions, so TWS decline is likely to continue, although perhaps at a slower rate, given that rainfall should normalize eventually and a 15% increase in rainfall is predicted by 2100.

Decreasing water storage in the Middle East has been quantified using GRACE by previous studies^{54–56}. Here we split the affected area into two regions, northwest Saudi Arabia (region 14; $-10.5 \pm 1.5 \text{ Gt yr}^{-1}$) and northern Middle East (region 15, which includes eastern Turkey, Syria, Iraq and Iran; $-32.1 \pm 1.5 \text{ Gt yr}^{-1}$). The declines result from a combination of recent drought and consequent increases in groundwater demand. Average precipitation during the study period was 78% and 96% of the 1979–2015 means in regions 14 and 15, respectively, with a slightly declining trend (-1% per year) in both. Although the irrigation dataset indicates that less than 1% of region 14 is irrigated, Landsat imagery reveals the appearance and expansion of crop irrigation over the past three decades, supplied by non-renewable groundwater. However, the Saudi Arabian government ended their domestic wheat production programme in market year 2014–15⁵⁷. Thus, although some farms have continued to operate, it is likely that the depletion rate in region 14 will diminish, and TWS may already be stabilizing (Extended Data Fig. 2).

Region 15 has experienced a more complicated recent water history^{54,58}. Turkey's construction of 22 dams upstream on the Tigris and Euphrates Rivers in the last three decades has considerably decreased the rate of flow into Iraq and Syria. Combined with long-term drought, this has forced widespread over-reliance on groundwater for both domestic and agricultural needs, and largely explains the large negative TWS trend^{54,59}. Surface and groundwater depletion is likely to continue in a stepwise fashion, with periods of near-stability during normal-to-wet years and rapid declines during drought years.

To the north, an adjoining zone of TWS depletion (region 16; $-18.1 \pm 1.3 \text{ Gt yr}^{-1}$) extends from the Ukraine through western Russia and into Kazakhstan. As before, the root cause of this depletion is competition for scarce water resources, exacerbated by drought. Fifteen per cent of the area is irrigated, including fertile croplands that are vital to Russia. Precipitation during the study period was 97% of normal, with a decreasing trend of 6 mm yr^{-1} (1% per year). As in region 15, surface- and groundwater depletion in region 16 is likely to continue as it has, stepwise, with substantial declines during drought years (2008, 2012 and 2014) and lesser recoveries in normal-to-wet years.

The water demands of regions 15 and 16 place severe pressure on the Aral and Caspian Seas⁶⁰ (regions 17 and 18, respectively). The demise of the Aral Sea is well known. Our estimate of the mass change in what remains of it is $-2.2 \pm 0.1 \text{ Gt yr}^{-1}$. Water level fluctuations in the Caspian Sea have previously been attributed to meteorological variability⁸ and direct evaporation from the sea⁶¹. We find that the annual discharge from the Volga River explains 60% of the variance in the annual mean level of the Caspian Sea compared with 18% explained by evaporation from the sea. Interannual variations in Volga River discharge are nearly three times as large as interannual variations in evaporation, and the former are controlled by both precipitation changes and the water demands of crops, which cover 37% of the basin. Using crop-production data and other information, we establish that the $-23.7 \pm 4.2 \text{ Gt yr}^{-1}$ change rate of the water mass in the Caspian Sea observed by GRACE was caused in part by diversions and direct withdrawals of water from the rivers that sustain it (see Methods for details), mirroring the circumstances that doomed the Aral Sea. Because the Caspian Sea contains about 78,000 Gt of water, at the current rate it will survive for three more millennia, but a receding shoreline could be an issue.

Three mass changes in Eurasia that are prominent in Fig. 1 are not associated with TWS at all. Crustal deformation accompanying the magnitude-9.1 Sumatra–Andaman earthquake of 2004 caused two of these mass changes, the dipole positive and negative trends in Sumatra and the Malay Peninsula, respectively⁶². The magnitude-9.0 Tohoku earthquake of 2011 caused the negative trend in Japan⁶³.

North America

Ongoing GIA processes centred near Hudson Bay, where the Laurentide ice sheet was thickest 20 to 95 thousand years ago, require a correction of the mass rates observed by GRACE of up to $5\text{--}6 \text{ cm yr}^{-1}$ (equivalent height of water)^{64,65}. However, GIA models are imperfect and thus there is large uncertainty in the apparent decreasing TWS trend in central Canada (region 19) and some evidence that it may reflect an overcorrection of GIA⁶⁶. Nevertheless, here we estimate the rate to be $-7.0 \pm 6.4 \text{ Gt yr}^{-1}$. Loss of water would be consistent with a recent study that concluded that Canada's subarctic lakes are vulnerable to drying when snow cover declines and that recent bouts of drying may be unprecedented in the past 200 years⁶⁷. On the other hand, precipitation was 102% of normal during the GRACE period, and a 17% increase is predicted by the end of the century.

The wetting trend in the northern Great Plains (region 20), $20.2 \pm 4.8 \text{ Gt yr}^{-1}$, arises from a combination of deep drought during 2001–2003, which depressed water levels greatly at the start of the GRACE period, followed by nine of the next eleven years having greater-than-average precipitation, including flooding in 2010–2011⁶⁸. The trend is likely to diminish over time, although a 7% increase in precipitation is predicted by 2100.

A historically severe drought centred in southern California (region 21) that began in 2007 (ignoring a wet 2010) and consequent

increases in groundwater demand^{69,70} conspired to diminish TWS at a rate of $-4.2 \pm 0.4 \text{ Gt yr}^{-1}$. Although atmospheric rivers replenished California's surface waters during 2016–2017 and policy changes have been enacted, it is doubtful that aquifer storage will recover completely without large usage reductions, in part because dewatering of aquifer materials can cause compaction of sediments, thus reducing aquifer capacity irrevocably⁷¹. In the Central Valley, which provides one-third of the vegetables and two-thirds of the fruits and nuts grown in the US, annual water demands for agriculture have exceeded renewable water resources since the early 20th century⁷¹. Groundwater well observations that extend back to 1962 suggest that each successive drought causes groundwater levels to step down to a new normal range without full recovery⁷¹, as in regions 15 and 16. Declining winter snowpack in the Sierra Nevada Mountains, including a 500-year low in 2015⁷², is a major concern because it is the main source of the region's surface-water supply and groundwater recharge.

Sporadic droughts⁷³ in region 22, which encompasses parts of the southern High Plains and Texas, produced an apparent trend of $-12.2 \pm 3.6 \text{ Gt yr}^{-1}$ during the GRACE period. In this case we forecast partial replenishment. Large precipitation variations caused TWS to seesaw between high and low (Extended Data Fig. 7). Heavy rains that led to flooding in parts of Texas and Oklahoma in May 2015, October 2015 and June 2016 ended the most recent drought and reduced the linear rate of TWS decline during the GRACE period. On the other hand, withdrawals of groundwater to support irrigated agriculture that exceed recharge in the central and southern High Plains aquifer have persisted for decades⁷⁴ and will continue until the resource is exhausted or management policies change. The fringes of the aquifer have already run dry in places, and recent estimates predict that the southern High Plains aquifer could be depleted within 30 years⁷⁴. Despite this situation, entrenched water rights are likely to preserve the status quo until the damage forces the hands of policymakers and stakeholders.

South America

Melting of the Patagonian ice fields (region 23) has previously been documented using altimetry⁷⁵ and GRACE⁷⁶. On the basis of our analysis (see Methods for details), TWS loss is occurring at a rate of $-25.7 \pm 5.1 \text{ Gt yr}^{-1}$. In a warming world, melting of the Patagonian ice fields will continue until they are exhausted.

The magnitude-8.8 Maule (Chile) earthquake that occurred on 27 February 2010 is partly responsible for the apparent trend in Central Argentina⁷⁷ (region 24). A model has not yet been developed to properly separate its effect from TWS variations after that date (Extended Data Fig. 3). TWS had previously been declining at a rate of $-8.6 \pm 1.2 \text{ Gt yr}^{-1}$. The region received substantially elevated precipitation in five of the six years between 1999 and 2004, producing a TWS surplus at the start of the GRACE period. Multi-year drought began in 2009, resulting in a negative trend observed from April 2002 to February 2010. TWS appears to have begun recovering (Extended Data Fig. 3) in response to above-normal precipitation in 2014 and 2015 (Extended Data Fig. 7), and we envisage that it will return to mean wetness conditions over time.

TWS increased during the GRACE period in central and western Brazil and its neighbours (region 25) at a rate of $51.9 \pm 9.4 \text{ Gt yr}^{-1}$. The region received less-than-average rainfall in every year from 2001 to 2005, followed by greater-than-average rainfall in six of the next ten years. As a result, TWS recovered from the early-period drought⁷⁸ and exhibited a massive, but transitory, increasing trend which may have already ended (Extended Data Fig. 3). The magnitude of this trend is explained by both the size of the region and the intensity of the Amazon water cycle⁷⁹. Still, we note that southern Brazil is a hotbed of dam construction⁴⁴, and it is possible that the filling of reservoirs contributed to the upward trend. Eastern Brazil (region 26) has recently suffered from a major drought⁸⁰, including well below normal rainfall in 2012, 2014 and 2015, causing TWS to plunge at a mean rate of $-16.7 \pm 2.9 \text{ Gt yr}^{-1}$ during the GRACE period. In both cases, assuming precipitation rates revert towards (or oscillate around) their long-term means, the

observed trends should fade. In fact, owing to the recent strong El Niño, 2015 was the driest year in the 37-year record for region 25 (Extended Data Fig. 3), which may portend a reversion to average TWS.

Africa

Six apparent trends stand out in Africa. In southern Africa, a powerful wetting trend, $29.5 \pm 3.5 \text{ Gt yr}^{-1}$, is observed in the western Zambezi basin, the Okavango delta and areas west of the coast (region 27). This region experienced a remarkable change in its hydroclimate. The area-averaged annual rainfall was less than 970 mm in every year from 1979 to 2005. That threshold was exceeded five times from 2006 to 2011. A permanent climatic shift was previously speculated on the basis of a significant decrease in annual precipitation in 1950–1975 and 1980–2005⁸¹. With ten years of additional hindsight, it appears that the region may have simply endured a prolonged drought from the late 1970s to the early 2000s. Thus, we attribute the GRACE-period trend to natural variability⁸². Although TWS appears to have peaked in 2012 (Extended Data Fig. 4), considering that the previous wet and dry periods lasted upwards of 25 years, it is plausible that the wetting trend could resume.

An apparent trend of $21.9 \pm 3.9 \text{ Gt yr}^{-1}$ occurs along the headwaters of the White Nile and Blue Nile rivers, including lakes Tanganyika and Victoria (region 28). Altimetry data indicate that during the study period both lakes experienced minimum water levels in 2006 and that their annual mean levels increased by 62 mm yr^{-1} and 40 mm yr^{-1} on average, respectively; these observations are consistent with the TWS time series. Together, the two lake level trends equate to less than a quarter (4.8 Gt yr^{-1}) of the observed TWS trend. Considering that, rainfall would seem to be the primary driver of TWS variations, while management of the large lakes⁸³ and dam building in the northern part of the region⁸⁴ also contribute. However, rainfall is not particularly well correlated with either TWS or lake levels. The lack of correlation may be indicative of inaccuracies stemming from the sparsity of rain gauges in the region. The observed rainfall trend was negligible during the study period, but a 12% increase is predicted by 2100. The northern part of region 28 encompasses the Grand Ethiopian Renaissance Dam on the Blue Nile River at Ethiopia's northwest border with Sudan, which Egypt has strongly denounced because of the possibility of reduced flow through the Nile. Construction of the dam began in 2011 and is ongoing. Filling of the 74-km^3 reservoir will probably produce a temporary increasing TWS trend in its immediate vicinity.

TWS has been increasing in tropical western Africa (region 29) at a rate of $24.1 \pm 2.1 \text{ Gt yr}^{-1}$. Precipitation was 3% below normal in 2000–2002 and 3% above normal during the rest of the GRACE period. This appears to be the primary cause of TWS accumulation, although the possible contribution of the many dams being built in this part of Africa⁴⁴ is unknown. Because interannual variability of rainfall is substantial in the region⁸⁵, disregarding the dams it is likely that the change rate of TWS will oscillate around zero over the coming decades. By 2100, rainfall is predicted to decrease by 6%; hence, the dam construction may be timely.

Decreasing TWS ($-7.2 \pm 1.0 \text{ Gt yr}^{-1}$) in region 30, which extends from the coast of central Africa into the northern Congo River basin, seems to be caused by natural interannual variability, although it has been suggested that the surface runoff rate has been enhanced by deforestation⁸⁴. Between 1999 and 2002 rainfall averaged 4% above normal, while it averaged 1% below normal during the rest of the GRACE period, including two very dry years in 2014 and 2015. The decrease in TWS is also consistent with the postulated negative correlation between TWS in the Amazon and Congo basins⁸⁶, which further implicates large-scale climatic oscillation as the ultimate driver⁸⁵.

The negative trend along the coast of southeastern Africa (region 31), $-12.9 \pm 2.3 \text{ Gt yr}^{-1}$, reflects a recent severe drought⁷⁹, which has caused major food shortages. Rainfall was 4% below average during the GRACE period, including annual accumulations that were below normal in five of the last eight years and barely above normal in the other three. Water levels in Lake Malawi, which is in the centre of the

region, are well correlated with regional TWS. The lake declined at a mean rate of 78 mm yr^{-1} during the period, accounting for 2.3 Gt yr^{-1} of the observed TWS trend. Hence, it is likely that the apparent trend is primarily caused by natural variability⁸⁴, although a 6% decrease in rainfall is predicted during this century.

A weak negative trend, $-11.7 \pm 2.9 \text{ Gt yr}^{-1}$, extends across arid Africa north of 19°N , excluding Morocco (region 32). The coefficient of determination is not large at 0.45; nevertheless, precipitation during the GRACE period was 7% above normal, which suggests that the consumptive use of fossil groundwater to stimulate agriculture and economic development is the cause^{55,84,87}. Three studies^{6,10,36} estimated recent rates of consumptive groundwater use across North Africa to be 7.8 Gt yr^{-1} , 15.7 Gt yr^{-1} and 4.1 Gt yr^{-1} , bracketing our TWS depletion estimate.

Australia

Australia appears to be bipolar with respect to water storage during the GRACE era, with wetting in the east and north and drying in the north-west. The worst drought in over 100 years afflicted eastern Australia during 2001–2009⁸⁸. It is likely that groundwater was more heavily consumed during that time to compensate for reduced availability of surface waters. Recovery from the drought began with heavy rains in 2010 and transitioned to severe flooding in 2011, with so much water stored on the continent in 2012 that the global mean sea level temporarily declined⁸⁹. The shift from dry to wet conditions caused the apparent wetting trend in region 33, $19.0 \pm 2.8 \text{ Gt yr}^{-1}$, but most of that water had already been shed by 2016 (Extended Data Fig. 4). Northern Western Australia received greater-than-normal rainfall during every year from 1997 to 2001, including the two wettest years in the GPCP record in 2000 and 2001. Thus, region 34 began 2002 near the maximum TWS capacity, and it gradually returned to average⁹⁰ ($-8.9 \pm 1.2 \text{ Gt yr}^{-1}$) with 99% of normal precipitation during the GRACE period. It is possible that aquifer dewatering associated with Pilbara's mining industry also contributed, but reliable data are not available to confirm and quantify that contribution. We can only justifiably conclude that natural variability is the primary explanation for both Australian trends.

Implications and discussion

GRACE has revealed considerable changes in freshwater resources occurring across the globe and has allowed them to be quantified at regional scales, unimpeded by sparse measurements or restrictive data-access policies. Some of these changes are manifestations of human water management that, before GRACE, were known only anecdotally, including TWS depletion in northern India, the North China Plain and the Middle East (regions 7, 12 and 14–16), or not at all, as in northwestern China (region 11). These changes portend a future in which already limited water resources will become even more precious. Others correlate well with global warming and predicted future precipitation changes, including worldwide ice-sheet and glacier melt (regions 1–4 and 23) and TWS increases in the northern high latitudes (regions 5–6). Apparent TWS trends in about one-third of the study regions represent partial cycles of longer-term interannual oscillations and may fade or reverse over the decades (see green dots in Fig. 1). Although we have made every effort to attribute the apparent trends properly, they will all require continued observation to better understand their causes and constrain their rates.

The GRACE data provide motivation for multilateral cooperation among nations, states and stakeholders, including development of trans-boundary water-sharing agreements, to balance competing demands and defuse potential conflict³³. Government policies that incentivize water conservation could help to avert a 'tragedy of the commons' scenario, that is, opportunistic competition for groundwater outweighing the altruistic impulse to preserve the resource. Northern India, the North China Plain, the Middle East and the area surrounding the Caspian Sea are already on a perilous path, while California, in response to severe drought and alarming groundwater declines in the Central Valley, recently passed legislation to regulate groundwater consumption.

In many regions, crop irrigation on massive scales has been supported by unsustainable rates of groundwater abstraction^{6,33–36,91}. In the face of aquifer depletion, population growth and climate change, water and food security will depend upon water-saving technologies and improved management and governance. The success of such an approach in arid Israel⁹² proves that a comprehensive water conservation strategy can work, and there are encouraging signs in Saudi Arabia (as previously discussed) and parts of India⁹³. Meanwhile, as China looks to improve living standards for its 1.38 billion residents, it will continue to face daunting water-management decisions, many of which are related to massive geoengineering and water-diversion projects that are likely to trigger political tensions.

The GRACE data also call attention to regions where continued monitoring will be essential for distinguishing, understanding and quantifying climate change impacts on the water cycle^{94,95} and groundwater^{96,97} in particular. This is important for two reasons. First, verification of emerging hydroclimatic trends, such as increasing northern high-latitude precipitation, would raise confidence in the ability of climate models to predict water-cycle consequences of climate change⁹⁸. Second, a redistribution of freshwater from dry to wet regions, as has been forecast, could exacerbate disparities between the water ‘haves’ and ‘have-nots’ and associated political instability, migration and conflict. Most groundwater depletion is occurring within Earth’s mid-latitudes, resulting in a positive drying feedback that is accelerating water losses and the severity of related socioeconomic issues³³.

New and future satellite remote-sensing missions that extend the long-term record of global hydrological observations will be essential for continued assessment of changing freshwater availability⁹⁹. In particular, the GRACE Follow On mission (planned to launch in early 2018), while affording a small increase in spatial resolution and accuracy¹⁰⁰, will enable surveillance of the trends described here and improved disentanglement of natural TWS variability from hydroclimatic change. Awareness of changing freshwater availability (for example, Fig. 1) is the first step towards addressing the challenges discussed here through improved infrastructure, water use efficiency, lifestyle and water-management decisions and policy.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0123-1>.

Received: 28 January 2017; Accepted: 12 March 2018;

Published online: 16 May 2018

- Changnon, S.A. *Detecting Drought Conditions in Illinois*. Circular 169 (Illinois State Water Survey, 1987).
- Rodell, M. & Famiglietti, J. S. An analysis of terrestrial water storage variations in Illinois with implications for the Gravity Recovery and Climate Experiment (GRACE). *Wat. Resour. Res.* **37**, 1327–1339 (2001).
- Getirana, A., Kumar, S., Giroto, M. & Rodell, M. Rivers and floodplains as key components of global terrestrial water storage variability. *Geophys. Res. Lett.* **44**, 10359–10368 (2017).
- Luthcke, S. B. et al. Antarctica, Greenland and Gulf of Alaska land ice evolution from an iterated GRACE global mascon solution. *J. Glaciol.* **59**, 613–631 (2013).
- Velicogna, I., Sutterley, T. C. & van den Broeke, M. R. Regional acceleration in ice mass loss from Greenland and Antarctica using GRACE time-variable gravity data. *Geophys. Res. Lett.* **41**, 8130–8137 (2014).
- Wada, Y., van Beek, L. P. H. & Bierkens, M. F. P. Unsustainable groundwater sustaining irrigation: a global assessment. *Wat. Resour. Res.* **48**, W00L06 (2012).
- Konikow, L. F. Contribution of global groundwater depletion since 1900 to sea-level rise. *Geophys. Res. Lett.* **38**, L17401 (2011).
- van Dijk, A. I. J. M., Renzullo, L. J., Wada, Y. & Tregoney, P. A global water cycle reanalysis (2003–2012) merging satellite gravimetry and altimetry observations with a hydrological multi-model ensemble. *Hydrol. Earth Syst. Sci.* **18**, 2955–2973 (2014).
- Zektser, I. S. & Everett, L. G. (eds) *Groundwater Resources of the World and Their Use* (UNESCO, Paris, 2004); <http://unesdoc.unesco.org/images/0013/001344/134433e.pdf>.
- Siebert, S. et al. Groundwater use for irrigation – a global inventory. *Hydrol. Earth Syst. Sci.* **14**, 1863–1880 (2010).
- Vörösmarty, C. J. et al. Global threats to human water security and river biodiversity. *Nature* **467**, 555–561 (2010).
- Syed, T. H., Famiglietti, J. S., Chambers, D. P., Willis, J. K. & Hilburn, K. Satellite-based global-ocean mass balance estimates of interannual variability and emerging trends in continental freshwater discharge. *Proc. Natl Acad. Sci. USA* **107**, 17916–17921 (2010).
- Rodell, M. et al. The observed state of the water cycle in the early 21st century. *J. Clim.* **28**, 8289–8318 (2015).
- Famiglietti, J. S. et al. Satellites provide the big picture. *Science* **349**, 684–685 (2015).
- Tapley, B. D., Bettadpur, S., Ries, J. C., Thompson, P. F. & Watkins, M. M. GRACE measurements of mass variability in the Earth system. *Science* **305**, 503–505 (2004).
- Wahr, J., Molenaar, M. & Bryan, F. Time variability of the Earth’s gravity field: hydrological and oceanic effects and their possible detection using GRACE. *J. Geophys. Res. Solid Earth* **103**, 30205–30229 (1998).
- Rodell, M. & Famiglietti, J. S. Detectability of variations in continental water storage from satellite observations of the time dependent gravity field. *Wat. Resour. Res.* **35**, 2705–2723 (1999).
- Swenson, S., Yeh, P. J. F., Wahr, J. & Famiglietti, J. A comparison of terrestrial water storage variations from GRACE with in situ measurements from Illinois. *Geophys. Res. Lett.* **33**, L16401 (2006).
- Cazenave, A. & Chen, J. Time-variable gravity from space and present-day mass redistribution in the Earth system. *Earth Planet. Sci. Lett.* **298**, 263–274 (2010).
- Rowlands, D. D. et al. Resolving mass flux at high spatial and temporal resolution using GRACE intersatellite measurements. *Geophys. Res. Lett.* **32**, L04310 (2005).
- Watkins, M. M., Wiese, D. N., Yuan, D. N., Boening, C. & Landerer, F. W. Improved methods for observing Earth’s time variable mass distribution with GRACE using spherical cap mascons. *J. Geophys. Res. Solid Earth* **120**, 2648–2671 (2015).
- Adler, R. et al. *The New Version 2.3 of the Global Precipitation Climatology Project (GPCP) Monthly Analysis Product* http://eagle1.umd.edu/GPCP_ICDR/GPCP_Monthly.html (2016).
- Salmon, J. M., Friedl, M. A., Froking, S., Wisser, D. & Douglas, E. M. Global rain-fed, irrigated, and paddy croplands: a new high resolution map derived from remote sensing, crop inventories and climate data. *Int. J. Appl. Earth Obs. Geoinf.* **38**, 321–334 (2015).
- Birkett, C., Reynolds, C., Beckley, B. & Doorn, B. in *Coastal altimetry* (eds Vignudelli, S. et al.) 19–50 (Springer, Berlin, 2011).
- Oldenborgh, G. J. et al. (eds) in *Climate Change 2013: The Physical Science Basis* (eds Stocker, T. F. et al.) 1311–1393 (Cambridge Univ. Press, Cambridge, 2013); http://www.ipcc.ch/pdf/assessment-report/ar5/wg1/WG1AR5_AnnexI_FINAL.pdf.
- Tamisiea, M. E., Leuliette, E. W., Davis, J. L. & Mitrovica, J. X. Constraining hydrological and cryospheric mass flux in southeastern Alaska using space-based gravity measurements. *Geophys. Res. Lett.* **32**, L20501 (2005).
- Gardner, A. S. et al. Sharply increased mass loss from glaciers and ice caps in the Canadian Arctic Archipelago. *Nature* **473**, 357–360 (2011).
- Boening, C., Lebedev, M., Landerer, F. & Stephens, G. Snowfall-driven mass change on the East Antarctic ice sheet. *Geophys. Res. Lett.* **39**, L21501 (2012).
- Schlegel, N.-J. et al. Application of GRACE to the assessment of model-based estimates of monthly Greenland Ice Sheet mass balance (2003–2012). *Cryosphere* **10**, 1965–1989 (2016).
- MacGregor, J. A. et al. Holocene deceleration of the Greenland Ice Sheet. *Science* **351**, 590–593 (2016).
- Reager, J. T. et al. A decade of sea level rise slowed by climate-driven hydrology. *Science* **351**, 699–703 (2016).
- Landerer, F. W., Dickey, J. O. & Güntner, A. Terrestrial water budget of the Eurasian pan-Arctic from GRACE satellite measurements during 2003–2009. *J. Geophys. Res. Atmos.* **115**, D23115 (2010).
- Famiglietti, J. S. The global groundwater crisis. *Nat. Clim. Chang.* **4**, 945–948 (2014).
- Gleeson, T., Wada, Y., Bierkens, M. F. & van Beek, L. P. Water balance of global aquifers revealed by groundwater footprint. *Nature* **488**, 197–200 (2012).
- Richey, A. S. et al. Uncertainty in global groundwater storage estimates in a total groundwater stress framework. *Wat. Resour. Res.* **51**, W198–W216 (2015).
- Döll, P., Schmied, H. M., Schuh, C., Portmann, F. T. & Eicker, A. Global-scale assessment of groundwater depletion and related groundwater abstractions: Combining hydrological modeling with information from well observations and GRACE satellites. *Wat. Resour. Res.* **50**, W698–W720 (2014).
- Long, D. et al. Global analysis of spatiotemporal variability in merged total water storage changes using multiple GRACE products and global hydrological models. *Remote Sens. Environ.* **192**, 198–216 (2017).
- Dalin, C., Wada, Y., Kastner, T. & Puma, M. J. Groundwater depletion embedded in international food trade. *Nature* **543**, 700–704 (2017).
- Phillips, T., Nerem, R., Fox-Kemper, B., Famiglietti, J. & Rajagopalan, B. The influence of ENSO on global terrestrial water storage using GRACE. *Geophys. Res. Lett.* **39**, L16705 (2012).
- Humphrey, V., Gudmundsson, L. & Seneviratne, S. I. Assessing global water storage variability from GRACE: trends, seasonal cycle, subseasonal anomalies and extremes. *Surv. Geophys.* **37**, 357–395 (2016).
- Rodell, M., Velicogna, I. & Famiglietti, J. S. Satellite-based estimates of groundwater depletion in India. *Nature* **460**, 999–1002 (2009).

42. Tiwari, V. M., Wahr, J. & Swenson, S. Dwindling groundwater resources in northern India, from satellite gravity observations. *Geophys. Res. Lett.* **36**, L18401 (2009).
43. Panda, D. K. & Wahr, J. Spatiotemporal evolution of water storage changes in India from the updated GRACE-derived gravity records. *Wat. Resour. Res.* **52**, 135–149 (2016).
44. Zarfl, C., Lumsdon, A. E., Berlekamp, J., Tydecks, L. & Tockner, K. A global boom in hydropower dam construction. *Aquat. Sci.* **77**, 161–170 (2015).
45. Wang, X., de Linage, C., Famiglietti, J. & Zender, C. S. Gravity Recovery and Climate Experiment (GRACE) detection of water storage changes in the Three Gorges Reservoir of China and comparison with in situ measurements. *Wat. Resour. Res.* **47**, W12502 (2011).
46. Chao, B. F., Wu, Y. H. & Li, Y. S. Impact of artificial reservoir water impoundment on global sea level. *Science* **320**, 212–214 (2008).
47. Zhang, G., Xie, H., Kang, S., Yi, D. & Ackley, S. F. Monitoring lake level changes on the Tibetan Plateau using ICESat altimetry data (2003–2009). *Remote Sens. Environ.* **115**, 1733–1742 (2011).
48. Zhang, T. Y. & Jin, S. G. Estimate of glacial isostatic adjustment uplift rate in the Tibetan Plateau from GRACE and GIA models. *J. Geodyn.* **72**, 59–66 (2013).
49. Jacob, T., Wahr, J., Pfeffer, W. T. & Swenson, S. Recent contributions of glaciers and ice caps to sea level rise. *Nature* **482**, 514–518 (2012).
50. Guo, M., Wu, W., Zhou, X., Chen, Y. & Li, J. Investigation of the dramatic changes in lake level of the Bosten Lake in northwestern China. *Theor. Appl. Climatol.* **119**, 341–351 (2015).
51. Stone, R. For China and Kazakhstan, no meeting of the minds on water. *Science* **337**, 405–407 (2012).
52. Hao, Y. et al. The role of climate and human influences in the dry-up of the Jinci Springs, China. *J. Am. Water Resour. Assoc.* **45**, 1228–1237 (2009).
53. Shamsudduha, M., Taylor, R. G. & Longuevergne, L. Monitoring groundwater storage changes in the highly seasonal humid tropics: validation of GRACE measurements in the Bengal Basin. *Wat. Resour. Res.* **48**, W02508 (2012).
54. Voss, K. A. et al. Groundwater depletion in the Middle East from GRACE with implications for transboundary water management in the Tigris-Euphrates-Western Iran region. *Wat. Resour. Res.* **49**, 904–914 (2013).
55. Sultan, M., Ahmed, M., Wahr, J., Yan, E. & Emil, M. in *Remote Sensing of the Terrestrial Water Cycle* (eds Lakshmi, V. et al.) 349–366 (John Wiley & Sons, Hoboken, 2014).
56. Joodaki, G., Wahr, J. & Swenson, S. Estimating the human contribution to groundwater depletion in the Middle East, from GRACE data, land surface models, and well observations. *Wat. Resour. Res.* **50**, 2679–2692 (2014).
57. USDA Foreign Agricultural Service. *Saudi Arabia Grain and Feed Annual, Global Agricultural Information Network*. Report number SA1602 (US Department of Agriculture, 2016); http://gain.fas.usda.gov/Recent%20GAIN%20Publications/Grain%20and%20Feed%20Annual/Riyadh_Saudi%20Arabia_3-14-2016.pdf.
58. Becker, R. H. The stalled recovery of the Iraqi marshes. *Remote Sens.* **6**, 1260–1274 (2014).
59. Chao, N., Luo, Z., Wang, Z. & Jin, T. Retrieving groundwater depletion and drought in the Tigris-Euphrates basin between 2003 and 2015. *Ground Water* (2017).
60. Zmijewski, K. & Becker, R. Estimating the effects of anthropogenic modification on water balance in the Aral Sea watershed using GRACE: 2003–12. *Earth Interact.* **18**, 1–16 (2014).
61. Chen, J. L. et al. Long-term Caspian Sea level change. *Geophys. Res. Lett.* **44**, 6993–7001 (2017).
62. Han, S.-C., Sauber, J., Luthcke, S. B., Ji, C. & Pollitz, S. S. Implications of postseismic gravity change following the great 2004 Sumatra-Andaman earthquake from the regional harmonic analysis of GRACE intersatellite tracking data. *J. Geophys. Res. Solid Earth* **113**, B11413 (2008).
63. Han, S. C., Sauber, J. & Riva, R. Contribution of satellite gravimetry to understanding seismic source processes of the 2011 Tohoku-Oki earthquake. *Geophys. Res. Lett.* **38**, L24312 (2011).
64. Peltier, W. R., Argus, D. F. & Drummond, R. Space geodesy constrains ice age terminal deglaciation: the global ICE-6G_C (VM5a) model. *J. Geophys. Res. Solid Earth* **120**, 450–487 (2015).
65. Peltier, W. R., Argus, D. F. & Drummond, R. Comment on “An Assessment of the ICE-6G_C (VM5a) glacial isostatic adjustment model by Purcell et al. *J. Geophys. Res. Solid Earth* **122**, 2019–2028 (2017).
66. Forman, B. A., Reichle, R. H. & Rodell, M. Assimilation of terrestrial water storage from GRACE in a snow-dominated basin. *Wat. Resour. Res.* **48**, W01507 (2012).
67. Bouchard, F. et al. Vulnerability of shallow subarctic lakes to evaporate and desiccate when snowmelt runoff is low. *Geophys. Res. Lett.* **40**, 6112–6117 (2013).
68. Reager, J. T. et al. Assimilation of GRACE terrestrial water storage observations into a land surface model for the assessment of regional flood potential. *Remote Sens.* **7**, 14663–14679 (2015).
69. Famiglietti, J. S. et al. Satellites measure recent rates of groundwater depletion in California's Central Valley. *Geophys. Res. Lett.* **38**, L03403 (2011).
70. Scanlon, B. R. et al. Groundwater depletion and sustainability of irrigation in the US High Plains and Central Valley. *Proc. Natl Acad. Sci. USA* **109**, 9320–9325 (2012).
71. Faunt, C. C., Sneed, M., Traum, J. & Brandt, J. T. Water availability and land subsidence in the Central Valley, California, USA. *Hydrogeol. J.* **24**, 675–684 (2016); erratum 25, 2215–2216 (2017).
72. Belmecheri, S., Babst, F., Wahl, E. R., Stahle, D. W. & Trouet, V. Multi-century evaluation of Sierra Nevada snowpack. *Nat. Clim. Chang.* **6**, 2–3 (2016).
73. Fernando, D. N. et al. What caused the spring intensification and winter demise of the 2011 drought over Texas? *Clim. Dyn.* **47**, 3077–3090 (2016).
74. Haack, E. M., Kendall, A. D. & Hyndman, D. W. Water level declines in the high plains aquifer: predevelopment to resource senescence. *Ground Water* **54**, 231–242 (2016).
75. Willis, M. J., Melkonian, A. K., Pritchard, M. E. & Ramage, J. M. Ice loss rates at the Northern Patagonian Icefield derived using a decade of satellite remote sensing. *Remote Sens. Environ.* **117**, 184–198 (2012).
76. Chen, J. L., Wilson, C. R., Tapley, B. D., Blankenship, D. D. & Ivins, E. R. Patagonia icefield melting observed by gravity recovery and climate experiment (GRACE). *Geophys. Res. Lett.* **34**, L22501 (2007).
77. Han, S. C., Sauber, J. & Luthcke, S. Regional gravity decrease after the 2010 Maule (Chile) earthquake indicates large-scale mass redistribution. *Geophys. Res. Lett.* **37**, L23307 (2010).
78. Chen, J. L., Wilson, C. R. & Tapley, B. D. The 2009 exceptional Amazon flood and interannual terrestrial water storage change observed by GRACE. *Wat. Resour. Res.* **46**, W12526 (2010).
79. Thomas, A. C., Reager, J. T., Famiglietti, J. S. & Rodell, M. A GRACE-based water storage deficit approach for hydrological drought characterization. *Geophys. Res. Lett.* **41**, 1537–1545 (2014).
80. Getirana, A. C. Extreme water deficit in Brazil detected from space. *J. Hydrometeorol.* **17**, 591–599 (2016).
81. Gaughan, A. E. & Waylen, P. R. Spatial and temporal precipitation variability in the Okavango-Kwando-Zambezi catchment, southern Africa. *J. Arid Environ.* **82**, 19–30 (2012).
82. Andersen, O. B. et al. in *Gravity, Geoid and Earth Observation, International Association of Geodesy Symposia* Vol. 135 (ed. Mertikas, S.) 521–526 (Springer, Berlin, 2010).
83. Swenson, S. & Wahr, J. Monitoring the water balance of Lake Victoria, East Africa, from space. *J. Hydrol.* **370**, 163–176 (2009).
84. Ahmed, M., Sultan, M., Wahr, J. & Yan, E. The use of GRACE data to monitor natural and anthropogenic induced variations in water availability across Africa. *Earth Sci. Rev.* **136**, 289–300 (2014).
85. Ndehedehe, C. E., Awange, J. L., Kuhn, M., Agutu, N. O. & Fukuda, Y. Climate teleconnections influence on West Africa's terrestrial water storage. *Hydrol. Processes* **31**, 3206–3224 (2017).
86. Crowley, J. W., Mitrovica, J. X., Bailey, R. C., Tamisiea, M. E. & Davis, J. L. Land water storage within the Congo Basin inferred from GRACE satellite gravity data. *Geophys. Res. Lett.* **33**, L19402 (2006).
87. Ramillien, G., Frappart, F. & Seoane, L. Application of the regional water mass variations from GRACE satellite gravimetry to large-scale water management in Africa. *Remote Sens.* **6**, 7379–7405 (2014).
88. van Dijk, A. J. M. et al. The Millennium Drought in southeast Australia (2001–2009): natural and human causes and implications for water resources, ecosystems, economy, and society. *Wat. Resour. Res.* **49**, 1040–1057 (2013).
89. Boening, C., Willis, J. K., Landerer, F. W., Nerem, R. S. & Fasullo, J. The 2011 La Niña: so strong, the oceans fell. *Geophys. Res. Lett.* **39**, L19602 (2012).
90. Munier, S., Becker, M., Maisongrande, P. & Cazenave, A. Using GRACE to detect groundwater storage variations: the cases of Canning Basin and Guarani aquifer system. *Int. Water Tech. J.* **2**, 2–13 (2012).
91. Jaramillo, F. & Destouni, G. Local flow regulation and irrigation raise global human water consumption and footprint. *Science* **350**, 1248–1251 (2015).
92. Fietelson, E. in *Water policy in Israel: Context, Issues and Options* (ed. Becker, N.) 15–32 (Springer Science & Business media, Dordrecht, 2013).
93. Bhanja, S. N. et al. Groundwater rejuvenation in parts of India influenced by water-policy change implementation. *Sci. Rep.* **7**, 7453 (2017).
94. Eicker, A., Forootan, E., Springer, A., Longuevergne, L. & Kusche, J. Does GRACE see the terrestrial water cycle “intensifying”? *J. Geophys. Res. Atmos.* **121**, 733–745 (2016).
95. Kusche, J., Eicker, A., Forootan, E., Springer, A. & Longuevergne, L. Mapping probabilities of extreme continental water storage changes from space gravimetry. *Geophys. Res. Lett.* **43**, 8026–8034 (2016).
96. Green, T. R. et al. Beneath the surface of global change: impacts of climate change on groundwater. *J. Hydrol.* **405**, 532–560 (2011).
97. Taylor, R. G. et al. Ground water and climate change. *Nat. Clim. Chang.* **3**, 322–329 (2013).
98. Swenson, S. C. & Milly, P. C. D. Climate model biases in seasonality of continental water storage revealed by satellite gravimetry. *Wat. Resour. Res.* **42**, W03201 (2006).
99. McCabe, M. F. et al. The future of Earth observation in hydrology. *Hydrol. Earth Syst. Sci.* **21**, 3879–3914 (2017).
100. Flechtner, F. et al. What can be expected from the GRACE-FO laser ranging interferometer for Earth science applications? *Surv. Geophys.* **37**, 453–470 (2016).

Acknowledgements We thank the German Space Operations Center of the German Aerospace Center (DLR) for providing nearly 100% of the raw telemetry data of the twin GRACE satellites. Landsat is an interagency programme managed by NASA and the US Geological Survey. Lake products are courtesy of the USDA/NASA G-REALM programme (available at http://www.pecad.fas.usda.gov/cropeplorer/global_reservoir/). V. Khan of the Hydrometeorological Research Center of the Russian Federation assisted with the Volga River discharge analysis. Graphics were produced by A. K. Moran, Global Science & Technology, Inc. This research was funded by NASA's GRACE Science Team and NASA's Energy and Water Cycle Study (NEWS) Team; the University of California Office of the President, Multicampus Research Programs and Initiatives; the

NASA Earth and Space Science Fellowship programme; the Jet Propulsion Laboratory; and the Ministry of Science and Technology, Taiwan. Portions of this research were conducted at the Jet Propulsion Laboratory, which is operated for NASA under contract with the California Institute of Technology.

Author contributions M.R. and J.S.F. performed background research and designed the study with input from J.T.R. and M.-H.L. D.N.W. and J.T.R. led the GRACE data and error analysis with assistance from F.W.L. M.R. and F.W.L. designed the figures with additional data prepared by H.K.B. M.R. and J.S.F. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0123-1>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.R.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

GRACE data have traditionally been processed by solving for gravity anomalies in terms of the Stokes coefficients (namely, C_{lm} and S_{lm} , with l denoting degree and m denoting order), which are the coefficients of the spherical harmonic expansion of Earth's gravity field^{16,101–106}. These solutions suffer from correlated errors that manifest as longitudinal striping in the gravity solution, which requires tailored 'destriping' and smoothing post-processing filters to remove¹⁰⁷. Although largely successful in removing errors, the post-processing also damps and smooths real geophysical signals¹⁰¹. Recent advances in GRACE data processing have shown that solving for gravity anomalies in terms of mass concentration (mascon) functions with carefully selected regularization results in superior localization of signals on an elliptical Earth^{4,21,108,109}. For instance, mascon solutions correlate better with in situ ocean-bottom pressure recorders than spherical-harmonic solutions^{21,109}, improve the spatial resolution of mass changes in Greenland²⁹ and were used to detect changes in the Atlantic meridional overturning circulation¹¹⁰. Currently, there are three publicly available GRACE mascon solutions: Jet Propulsion Laboratory mascons RL05M.1 version 2^{21,111} (JPL-M), Center for Space Research mascons RL05M¹⁰⁹ (CSR-M) and Goddard Space Flight Center mascons version 2.3b⁴ (GSFC-M). JPL-M parameterizes the gravity field with 4,551 equal-area 3° mascon elements, whereas CSR-M and GSFC-M both parameterize the gravity field in terms of 1° mascon elements (~41,000 mascon elements are solved for in each solution). Although the implementation details of each mascon solution differ, we note that the JPL-M solution has the unique characteristic that each 3° mascon element is relatively uncorrelated with neighbouring mascon elements, whereas the 1° mascon elements in CSR-M and GSFC-M solutions are highly correlated with their neighbours. Three degrees correspond approximately to the 'native' resolution of GRACE, and the lack of correlation between neighbouring mascon elements in the retrieval allows for a quantitative understanding of leakage errors when aggregating mass anomalies within a hydrological basin¹¹²—in fact, no literature yet exists on quantifying leakage errors in 1° mascon solutions. Therefore, in this work we use JPL-M for trend analysis and mapping; however, we use all three mascon solutions (JPL-M, CSR-M and GSFC-M) to derive uncertainties.

The JPL-M solution parameterizes each monthly gravity field in terms of 4,551 equal-area, surface spherical-cap mass-concentration functions and uses a regularization approach that implements both spatial and temporal correlations to remove correlated errors during the gravity inversion. A coastline resolution improvement filter is used to separate between land and ocean mass within mascons that span coastlines¹¹². Because GRACE does not produce a reliable estimate of Earth's oblateness (C_{20} coefficient), we follow the standard protocol of using satellite laser ranging to provide this estimate¹¹³. Further, GRACE gravity-field anomalies are measured in the centre-of-mass reference frame of Earth and therefore need to be augmented with a 'geocentre' estimate to capture all surface-mass changes¹¹⁴. GIA corrections are made using the updated ICE-6G_D model^{64,65}, with an exception for Antarctica, for which we reduce the fitted rate of mass change by 9.2 Gt yr^{-1} on the basis of a regional model¹¹⁵ that potentially provides a better GIA estimate for Antarctica¹¹⁶. Finally, corrections are made to the C_{21} and S_{21} coefficients¹¹⁷ (degree-2 coefficients are related to the moments and products of inertia with respect to a defined reference frame and set of background force models) to fully remove the pole tide from the GRACE data. Jumps in the background atmosphere and ocean dealiasing product are corrected as well¹¹⁸.

Prior to computing the best-fit linear trend from a TWS time series, the seasonal cycle was removed as follows. First, missing months of data were filled by linear interpolation. Next, the mean monthly seasonal cycle was computed by averaging all Januaries, all Februaries, etc. Finally, for each month in the original, non-gap-filled time series, the mean for the corresponding month of the year was subtracted. The first step (gap filling) was necessary because, for example, the month of May was under-sampled in the second half of the study period, which caused the mean May to be biased in locations where a consistent trend existed (that is, most of the regions of this study).

Trend error estimates account for both systematic and random GRACE measurement errors, as well as the systematic error of the GIA model. The GRACE measurement error is taken to be 1σ , where σ is the standard deviation between trend estimates obtained from JPL-M, CSR-M and GSFC-M. Given the specific basin boundaries used in this study, we find JPL-M to have more pronounced trends (both positive and negative) than CSR-M and GSFC-M, which is consistent with previous conclusions¹¹⁹. This spread is due to a fundamental difference in the spectral content between the 3° mascons and 1° mascons, implying that leakage characteristics are different when aggregating mass anomalies over a particular region (somewhat counter-intuitively, the 3° mascons 'focus' more signals than the 1° sampled mascons). In essence, the 'smooth' nature of the 1° mascon solutions (CSR-M and GSFC-M) results in considerable damping of the signal over our regions of interest owing to leakage across the basin boundaries. For a more direct comparison of the three solutions over our regions of interest, we matched the spectral content of JPL-M to that of CSR-M. The regularization of the CSR

mascon solution is based on a smoothed (using a 200-km Gaussian) representation of a regularized spherical-harmonic solution¹⁰⁹. Hence, it is expected that the final mascon solution will inherit some of these spectral characteristics. Therefore, we smooth JPL-M with a 200-km-radius Gaussian filter and compare the trend estimates of the smoothed version of JPL-M to those of CSR-M and GSFC-M. Thus, the agreement is substantially improved, and trends in the smoothed version of JPL-M are also damped similarly to CSR-M and GSFC-M (see Extended Data Fig. 9 for an example). Similar analysis has been performed before in a study of mass variations over the Caspian Sea¹²⁰. We use the standard deviation of trend estimates obtained from the smoothed version of JPL-M, CSR-M and GSFC-M to derive the GRACE measurement errors. The GIA model error is taken to be the 1σ spread between four competing GIA models^{64,65,121–124} that implement two distinct loading histories, four distinct viscosity profiles and different implementations of physics. The uncertainty on the trend for any region is given by the root sum of squares combining the GIA model error (which manifests only as a trend) and the GRACE measurement error.

Time series for the Aral and Caspian seas (regions 17 and 18) were calculated by applying a set of gain factors to the GRACE data. Gain factors redistribute mass within each individual mascon (at sub-mascon resolution), allowing exact averaging kernels to be applied to a region of interest and retrieval of accurate, unbiased (by leakage) mass-change values^{101,112}. These particular gain factors were derived¹¹² using a combination of total-column soil moisture output from the Noah land-surface model driven by the Global Land Data Assimilation System¹²⁵ (which does not include sea-water variations) along with altimetry data¹²⁶ over the Aral and Caspian seas.

Recent variations in Caspian Sea level have been attributed by previous studies to natural meteorological variability⁸ and direct evaporation from the sea surface⁶¹. We tested these two theories as well as a third, agricultural water consumption. Flow in the Volga River, which delivers roughly 80% of the runoff to the Caspian Sea, is controlled by a series of eleven dams¹²⁷. Among other purposes, these ensure a steady supply of water for crop irrigation¹²⁷. No data were available to quantify interannual variations in irrigation extent, intensity or volumes in the Caspian Sea drainage basin during the study period. Estimates of Russian annual wheat, maize, rice and soybean production¹²⁸ (in tonnes) during 1992–2015 were obtained from the Organisation for Economic Co-operation and Development (OECD). According to the irrigation dataset²³, the Volga River basin, which drains to the Caspian Sea, includes 3% irrigated crops and 37% rain-fed crops by area, and it accounts for about half of all Russian crop production. Therefore, Russian crop production is a fair, but imperfect, indicator of agricultural water demand in the basin. Yearly total production was normalized by subtracting the 24-year mean and dividing it by the standard deviation. Normalization was similarly performed on the annual time series of GPCP precipitation²² over the Caspian Sea and Volga River drainage basins, the Volga River discharge, reanalysis-based Caspian Sea evaporation¹²⁹ and changes in Caspian Sea level obtained from satellite altimetry²⁴. Correlation coefficients (and significance levels) between normalized Caspian Sea level change and its significant drivers (Extended Data Fig. 10) were 0.78 (Volga River discharge; $P < 0.001$), -0.47 (crop production; $P = 0.02$), -0.43 (Caspian Sea evaporation; $P = 0.04$) and 0.41 (Caspian Sea drainage basin precipitation; $P = 0.05$). Correlation coefficients (and significance levels) between normalized Volga River discharge and significant drivers were 0.52 (Volga River basin precipitation; $P = 0.01$) and -0.40 (crop production; $P = 0.06$). Notably, the correlation between crop production and precipitation was negligible, suggesting that irrigation effectively mitigates the impact of drought. Interannual variations in Caspian Sea evaporation do indeed contribute significantly to Caspian Sea level changes. However, annual Volga River discharge variations are better correlated with annual changes in Caspian Sea level, they are larger than variations in Caspian Sea evaporation (standard deviation of 48 Gt versus 18 Gt, compared with a mean magnitude of annual Caspian Sea level change of 38 Gt) and they are controlled by both precipitation and rising agricultural water demand¹²⁷. We therefore conclude that all three factors contributed to the observed water loss ($-23.7 \pm 4.2 \text{ Gt yr}^{-1}$ from GRACE, ignoring steric effects; -25.4 Gt yr^{-1} from satellite altimetry) during 2002–2015.

For the Gulf of Alaska coast and the Patagonian ice fields (regions 3 and 23), it was also necessary to increase the rates of mass loss (by 7 Gt yr^{-1} and 9 Gt yr^{-1} , respectively) to account for Little Ice Age GIA³¹. We note that the full GIA corrections to Antarctica, the Gulf of Alaska coast and the Patagonian ice fields are not incorporated into Extended Data Fig. 1 and 3.

The irrigated area fractions (Table 1) were computed by area-weighted averaging of the individual pixel values of irrigation intensity²³ (%) within each study region. Precipitation trends (mm yr^{-1}) were computed on the basis of monthly data²², as above for TWS, except that there were no gaps to fill. Precipitation trends (% per year) and percentages of normal precipitation were computed using the 1979–2015 annual mean precipitation totals for each region. Predicted precipitation changes were computed as area-weighted averages from the IPCC dataset²⁵ over the study

regions. The precipitation maps in Fig. 2 were computed as above, but on a pixel-by-pixel basis.

The explanation for the mass-loss trend in northwestern China (region 11), $-5.5 \pm 0.5 \text{ Gt yr}^{-1}$, is complex. Drought was not a factor, given that precipitation was 10% above normal and stable during the period. Two recent studies^{130,131} estimated the rate of glacier loss over the entire Tien Shan mountain range to be $-5.4 \pm 2.9 \text{ Gt yr}^{-1}$ and $-7.5 \pm 3.4 \text{ Gt yr}^{-1}$ based on Ice, Cloud and Land Elevation Satellite (ICESat) observations from 2003 to 2009. These estimates are somewhat smaller than our GRACE-based estimate of TWS decline in region 11 during that period ($-8.3 \pm 0.8 \text{ Gt yr}^{-1}$), despite region 11 encompassing less than half of the area of glacier melt¹³⁰. Hence, we conjecture that an additional catalyst for mass loss must exist. Xinjiang province is one of the world's largest producers of coal, having an estimated 2.2 trillion tons of reserves¹³². Reported rates of coal removal and burning are more than an order of magnitude smaller than the GRACE-observed mass loss¹³², but mining involves dewatering of the aquifers that the mines intersect. Consequent groundwater depletion in the area is possible⁵² but unconfirmed. Adding to the complexity, region 11 lies within a larger endorheic basin, which means that water pumped from the ground or melting from glaciers will remain as surface water, become groundwater recharge or evapotranspire, as opposed to flowing to the ocean. However, on the basis of satellite altimetry data, the elevations of the five lakes within the surrounding endorheic basin did not increase during the study period. All either declined or did not change significantly. The two lowlands into which region 11 drains (one to the northwest, one to the southeast) have GRACE-based trends of 0.3 Gt yr^{-1} and -0.6 Gt yr^{-1} (both insignificant). Ultimately, evapotranspiration must account for the water lost from region 11. The average annual precipitation in region 11 is 194 mm yr^{-1} , making it the fourth-driest of the 32 study regions. The endorheic basin is extensively irrigated, including 7% of region 11, and irrigation intensity is likely rising in support of Xinjiang province's population growth (from 18.2 million in 2000 to 21.8 million in 2010)⁵¹. Massive amounts of surface water from Lake Bosten and the Kongque River (both to the southeast of region 11) are transferred via aqueducts southwards to the Tarim River to support farming in the arid plains; however, the Tarim River runs dry before reaching its natural terminus, Lop Nor lake⁵⁰. To summarize, the Tien Shan mountain glaciers in region 11 are shrinking because of global warming. Groundwater may be declining owing to agricultural withdrawals or mining operations, but the latter is unconfirmed. Because region 11 lies within an endorheic basin, neither glacier melt nor groundwater pumping can alone explain the observed TWS depletion. The corollary is that the resulting additions to surface water are balanced by desert- and irrigation-enhanced evapotranspiration.

As noted in the main text, although previous GRACE-based studies of the North China Plain (region 12) agreed that groundwater depletion associated with intense irrigation was the cause of the trend, they offered a wide range of estimates of the TWS or groundwater trend. Specifically, these estimates were -8.3 Gt yr^{-1} over a $370,000\text{-km}^2$ area¹³³, -35 Gt yr^{-1} over a $2,086,000\text{-km}^2$ area¹³⁴, -2.33 Gt yr^{-1} over a $370,000\text{-km}^2$ area¹³⁵ and $-14.09 \text{ Gt yr}^{-1}$ over a $1,500,000\text{-km}^2$ area¹³⁶, compared with our estimate of -11.3 Gt yr^{-1} over an $876,004\text{-km}^2$ area.

Data availability. Specific sources of data used in this study were the following. The primary GRACE TWS dataset is JPL Mascon RL05M.1 version 2, accessed on 3 February 2017 from https://grace.jpl.nasa.gov/data/get-data/jpl_global_mascons/. Additional GRACE TWS datasets used to estimate errors were CSR RL05 Mascon version 1, accessed on 20 September 2017 from http://www2.csr.utexas.edu/grace/RL05_mascons.html, and GSFC Mascon version 2.3b, accessed on 5 October 2017 from <https://neptune.gsfc.nasa.gov/gnophysics/index.php?section=413>. Primary GIA data used in this study were the ICE-6G model, accessed on 1 December 2017 from <http://www.atmos.physics.utoronto.ca/~peltier/data.php>, and the IJ05_R2 GIA correction for Antarctica, accessed on 3 February 2018 from <http://onlinelibrary.wiley.com/doi/10.1002/jgrb.50208/full>. Additional GIA data used to compute the GIA model error included ICE-6G_ANU_D, accessed on 3 February 2018 from <http://onlinelibrary.wiley.com/doi/10.1002/2017JB014930/full>, the A et al. (2013)¹²¹ GIA model, accessed on 16 December 2013 from <ftp://podaac-ftp.jpl.nasa.gov/allData/tellus/L3/pgr/>, and the Paulson et al. (2007)¹²² GIA model, accessed on 3 February 2018 from <https://academic.oup.com/gji/article/171/2/497/2018541>. Atmosphere and ocean dealiasing product jump corrections were accessed on 13 June 2016 from <ftp://podaac-ftp.jpl.nasa.gov/allData/grace/docs/>. Precipitation data from GPCP version 2.3 were accessed on 23 September 2016 from <https://www.esrl.noaa.gov/psd/data/gridded/data.gpcp.html>. Global rain-fed, irrigated and paddy croplands version-1 data were accessed on 12 September 2016 from <http://ftp-earth.bu.edu/public/friedl/GRIPCmap/>. Global reservoir/lake elevation TPJO.2.3 data were accessed on 29 July 2016 from https://ipad.fas.usda.gov/cropeplorer/global_reservoir/. Precipitation change data predicted by the IPCC 5th Assessment Report (RCP8.5) were accessed on 1 September 2016 from https://www.ipcc.ch/pdf/assessment-report/ar5/wg1/WG1AR5_AnnexI_FINAL.pdf. Russian crop production data were accessed on 16 August

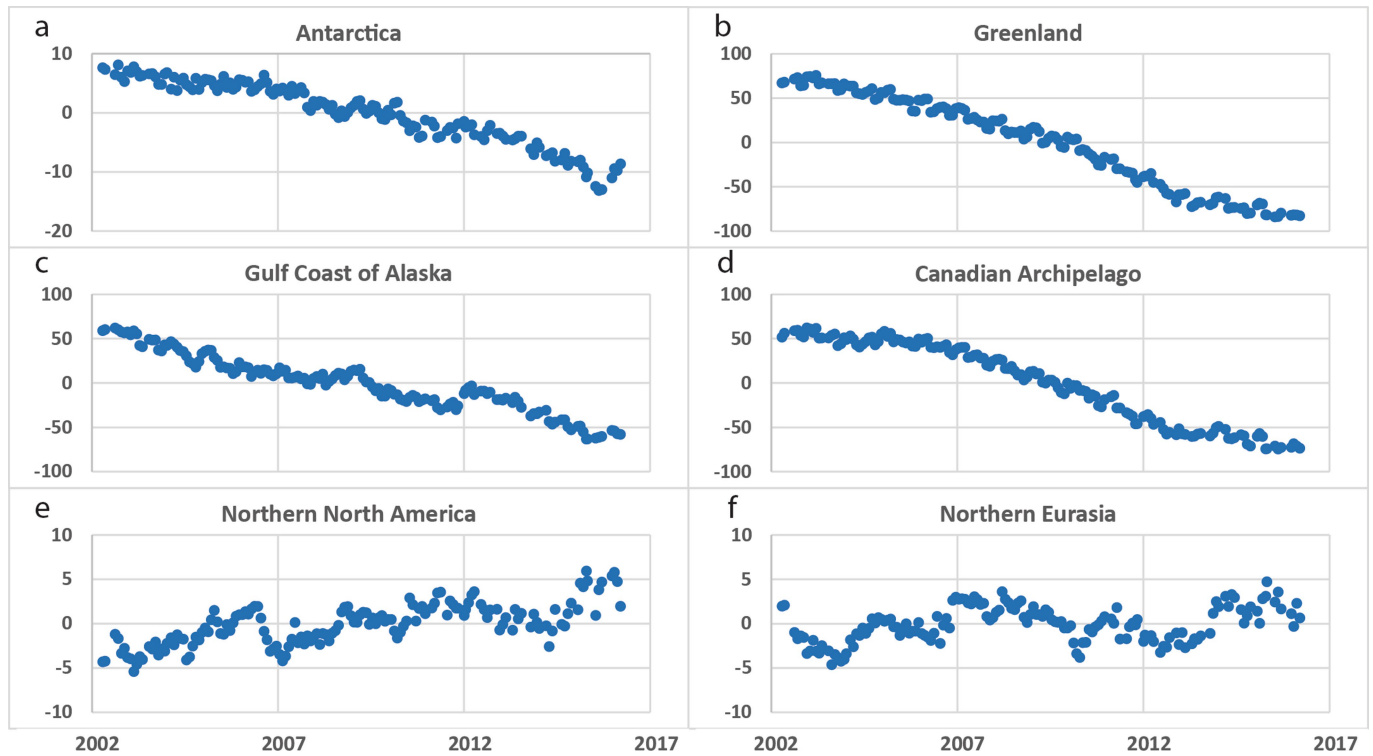
2017 from <https://data.oecd.org/agroutput/crop-production.htm>. Latent heat flux (evapotranspiration) data for the Caspian Sea and its drainage basin were extracted from MERRA-2 version M2TMNXLND_5.12.4, accessed on 19 September 2017 from https://disc.sci.gsfc.nasa.gov/datasets/M2TMNXLND_5.12.4/summary. Volga River discharge observations are restricted from public access, but a time series of normalized annual discharge values was provided to M.R. by V. Khan of the Hydrometeorological Research Center of the Russian Federation.

The JPL RL05M GRACE solution used in this study is identical to that available from the NASA/JPL GRACE Tellus website, with the exception that we implemented a different GIA model, a correction to the pole tide and corrections to the background atmosphere and ocean dealiasing model. These adjustments are available from D.N.W. upon request. Data analysed to create Extended Data Fig. 9 are available from D.N.R. upon request. Excel spreadsheets containing the data and calculations used to create Table 1 and Extended Data Fig. 10 are available from M.R. upon request.

Code availability. MATLAB scripts were used to prepare GRACE-based TWS time series for the study regions, including GIA adjustments, C_{21} and S_{21} coefficient replacements, and corrections for jumps in the atmosphere and ocean dealiasing products. These are available from D.N.W. upon reasonable request. TWS time series analyses, including trend estimation and r^2 computation, were performed within Excel spreadsheets, which are available from M.R. upon reasonable request.

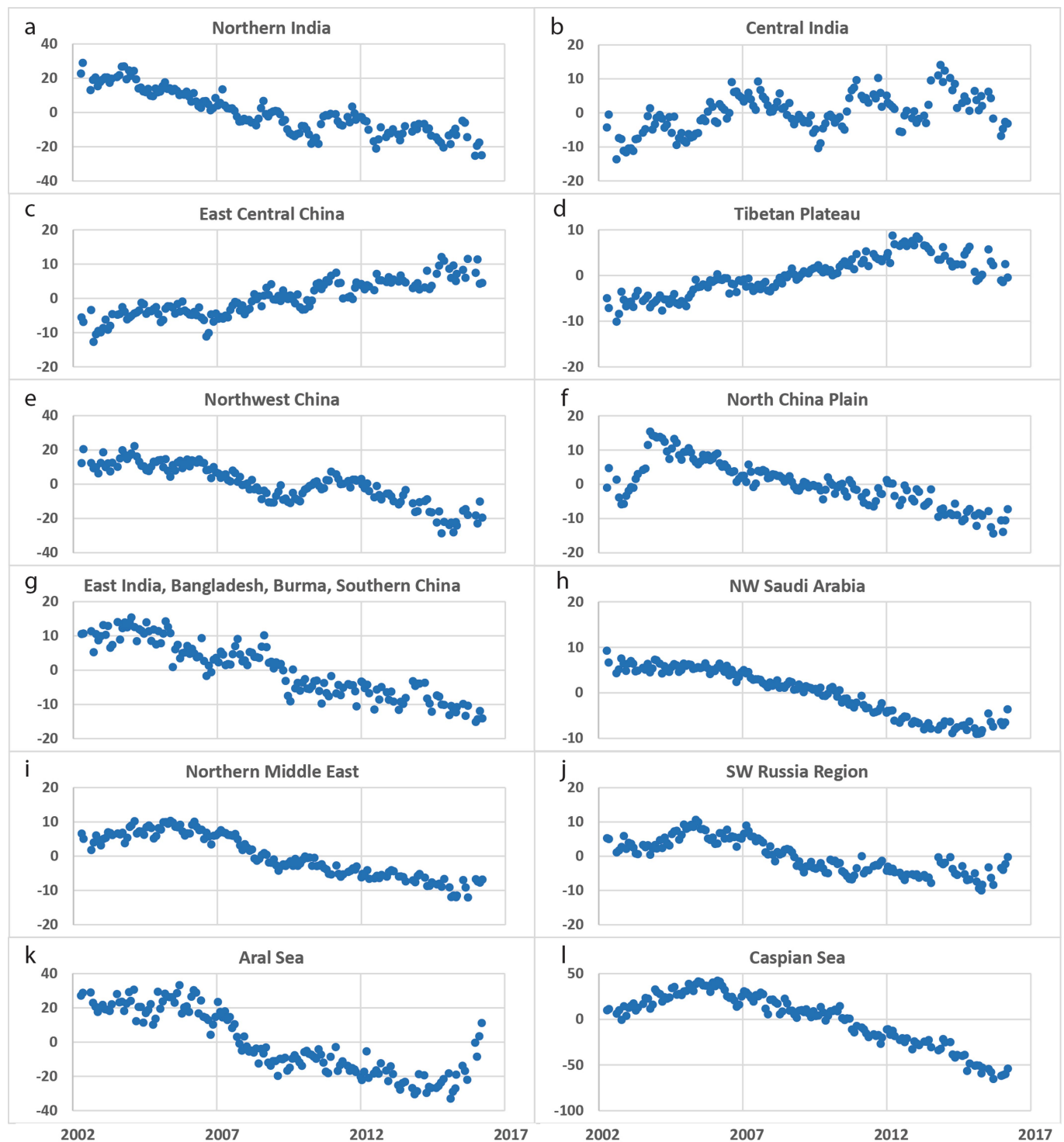
101. Landerer, F. W. & Swenson, S. C. Accuracy of scaled GRACE terrestrial water storage estimates. *Wat. Resour. Res.* **48**, W04531 (2012).
102. Dahle, C. et al. in *Observation of the System Earth from Space-CHAMP, GRACE, GOCE and Future Missions* (eds Flechtner, F. et al.) 29–39 (Springer, Berlin, 2014).
103. Mayer-Gürr, T. et al. *ITSG-Grace2016 - Monthly and Daily Gravity Field Solutions from GRACE* <https://doi.org/10.5880/igcm.2016.007> (2016).
104. Bruinsma, S., Lemoine, J.-M., Biancale, R. & Vales, N. CNES/GRGS 10-day gravity field models (release 02) and their evaluation. *Adv. Space Res.* **45**, 587–601 (2010).
105. Kurtenbach, E. et al. Improved daily GRACE gravity field solutions using a Kalman smoother. *J. Geodyn.* **59–60**, 39–48 (2012).
106. Liu, X. et al. DEOS Mass Transport model (DMT-1) based on GRACE satellite data: methodology and validation. *Geophys. J. Int.* **181**, 769–788 (2010).
107. Swenson, S. & Wahr, J. Post-processing removal of correlated errors in GRACE data. *Geophys. Res. Lett.* **33**, L08402 (2006).
108. Andrews, S. B., Moore, P. & King, M. A. Mass change from GRACE: a simulated comparison of Level-1B analysis techniques. *Geophys. J. Int.* **200**, 503–518 (2014).
109. Save, H., Bettadpur, S. & Tapley, B. D. High resolution CSR GRACE RL05 mascons. *J. Geophys. Res. Solid Earth* **121**, 7547–7569 (2016).
110. Landerer, F. W., Wiese, D. N., Bentel, K., Boening, C. & Watkins, M. M. North Atlantic meridional overturning circulation variations from GRACE ocean bottom pressure anomalies. *Geophys. Res. Lett.* **42**, 8114–8121 (2015).
111. Wiese, D. N., Yuan, D.-N., Boening, C., Landerer, F. W. & Watkins, M. M. *JPL GRACE Mascon Ocean, Ice, and Hydrology Equivalent Water Height RL05M.1 CRI Filtered Version 2 PO.DAAC, CA, USA* <https://doi.org/10.5067/TEMSC-2LCR5> (2016).
112. Wiese, D. N., Landerer, F. W. & Watkins, M. M. Quantifying and reducing leakage errors in the JPL RL05M GRACE mascon solution. *Wat. Resour. Res.* **52**, 7490–7502 (2016).
113. Cheng, M. & Tapley, B. D. Variations in the Earth's oblateness during the past 28 years. *J. Geophys. Res.* **109**, B09402 (2004).
114. Swenson, S., Chambers, D. & Wahr, J. Estimating geocenter variations from a combination of GRACE and ocean model output. *J. Geophys. Res.* **113**, B08410 (2008).
115. Ivins, E. R. et al. Antarctic contribution to sea level rise observed by GRACE with improved GIA correction. *J. Geophys. Res.* **118**, 3126–3141 (2013).
116. Shepherd, A. et al. A reconciled estimate of ice-sheet mass balance. *Science* **338**, 1183–1189 (2012).
117. Wahr, J., Nerem, R. S. & Bettadpur, S. V. The pole tide and its effect on GRACE time variable gravity measurements: Implications for estimates of surface mass variations. *J. Geophys. Res. Solid Earth* **120**, 4597–4615 (2015).
118. Fagioli, E., Flechtner, F., Horwath, M. & Döbslaw, H. Correction of inconsistencies in ECMWF's operational analysis data during de-aliasing of GRACE gravity models. *Geophys. J. Int.* **202**, 2150–2158 (2015).
119. Scanlon, B. R. et al. Global evaluation of new GRACE mascon products for hydrologic applications. *Water Resour. Res.* **52**, 9412–9429 (2016).
120. Chen, J. L., Wilson, C. R., Tapley, B. D., Save, H. & Cretaux, J.-F. Long-term and seasonal Caspian Sea level change from satellite gravity and altimeter measurements. *J. Geophys. Res. Solid Earth* **122**, 2274–2290 (2017).
121. A, G., Wahr, J. & Zhong, S. Computations of the viscoelastic response of a 3-D compressible Earth to surface loading: an application to Glacial Isostatic Adjustment in Antarctica and Canada. *Geophys. J. Int.* **192**, 557–572 (2013).
122. Paulson, A., Zhong, S. & Wahr, J. Inference of mantle viscosity from GRACE and relative sea level data. *Geophys. J. Int.* **171**, 497–508 (2007).
123. Purcell, A., Tregoning, P. & Dehecq, A. An assessment of the ICE6G_C(VM5a) glacial isostatic adjustment model. *J. Geophys. Res. Solid Earth* **121**, 3939–3950 (2016).

124. Purcell, A., Tregoning, P. & Dehecq, A. Reply to comment by W. R. Peltier, D. F. Argus, and R. Drummond on "An assessment of the ICE6G_C (VM5a) glacial isostatic adjustment model". *J. Geophys. Res. Solid Earth* **123**, 2029–2032 (2017).
125. Rodell, M. et al. The global land data assimilation system. *Bull. Am. Meteorol. Soc.* **85**, 381–394 (2004).
126. Crétaux, J.-F. et al. SOLS: a lake database to monitor in the near real time water level and storage variations from remote sensing data. *Adv. Space Res.* **47**, 1497–1507 (2011).
127. Avakyan, A. B. Volga-Kama cascade reservoirs and their optimal use. *Lakes Reservoirs: Res. Manage.* **3**, 113–121 (1998).
128. OECD. *Crop Production (Indicator)* <https://data.oecd.org/agroutput/crop-production.htm> (2017).
129. Gelaro, R. et al. The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *J. Clim.* **30**, 5419–5454 (2017).
130. Farinotti, D. et al. Substantial glacier mass loss in the Tien Shan over the past 50 years. *Nat. Geosci.* **8**, 716–722 (2015).
131. Gardner, A. et al. A reconciled estimate of glacier contributions to sea level rise: 2003 to 2009. *Science* **340**, 852–857 (2013).
132. Mou, D. & Li, Z. A spatial analysis of China's coal flow. *Energy Policy* **48**, 358–368 (2012).
133. Feng, W. et al. Evaluation of groundwater depletion in North China using the Gravity Recovery and Climate Experiment (GRACE) data and ground-based measurements. *Wat. Resour. Res.* **49**, 2110–2118 (2013).
134. Moiwo, J. P., Tao, F. & Lu, W. Analysis of satellite-based and in situ hydro-climatic data depicts water storage depletion in North China Region. *Hydrol. Processes* **27**, 1011–1020 (2013).
135. Tang, Q., Zhang, X. & Tang, Y. Anthropogenic impacts on mass change in North China. *Geophys. Res. Lett.* **40**, 3924–3928 (2013).
136. Ebead, B., Ahmed, M., Niu, Z. & Huang, N. Quantifying the anthropogenic impact on groundwater resources of North China using Gravity Recovery and Climate Experiment data and land surface models. *J. Appl. Remote Sens.* **11**, 026029 (2017).

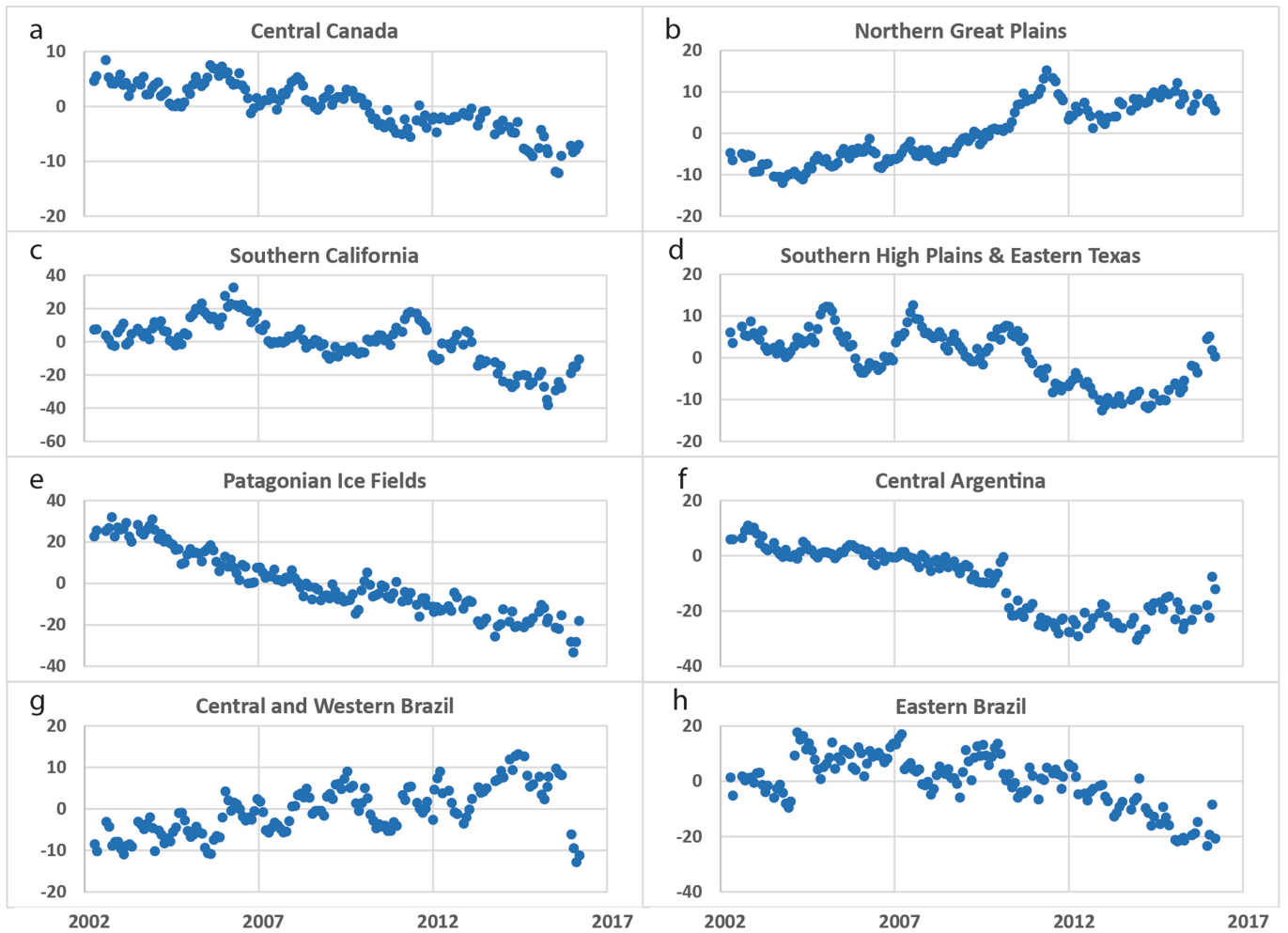


Extended Data Fig. 1 | Non-seasonal TWS anomalies—global regions. a–f, Time series of monthly TWS anomalies (departures from the period mean) from GRACE, after removing the mean seasonal cycle, averaged

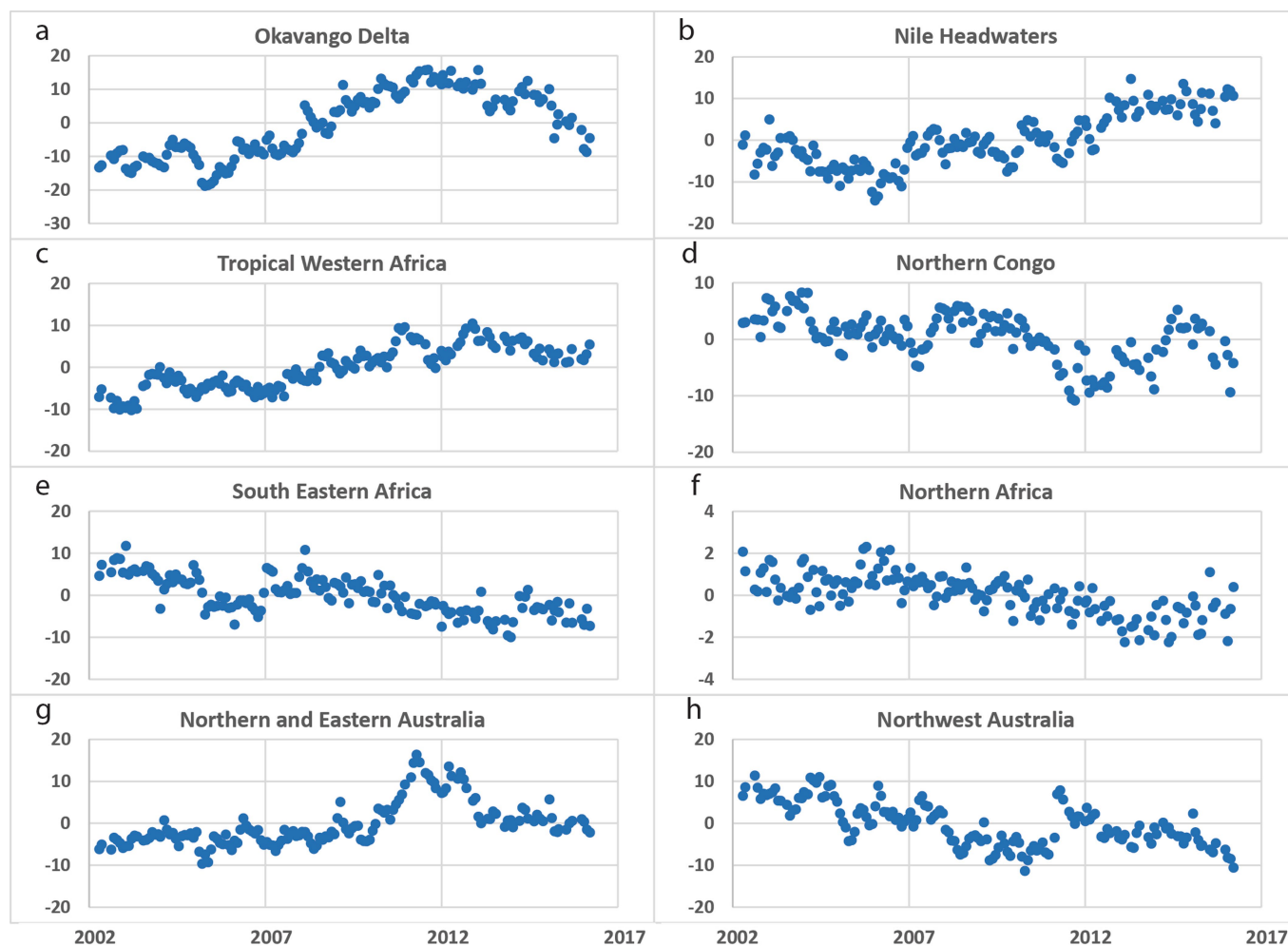
over each of study regions 1–6, expressed as equivalent heights of liquid water (in centimetres). We note that the y axes vary among panels.



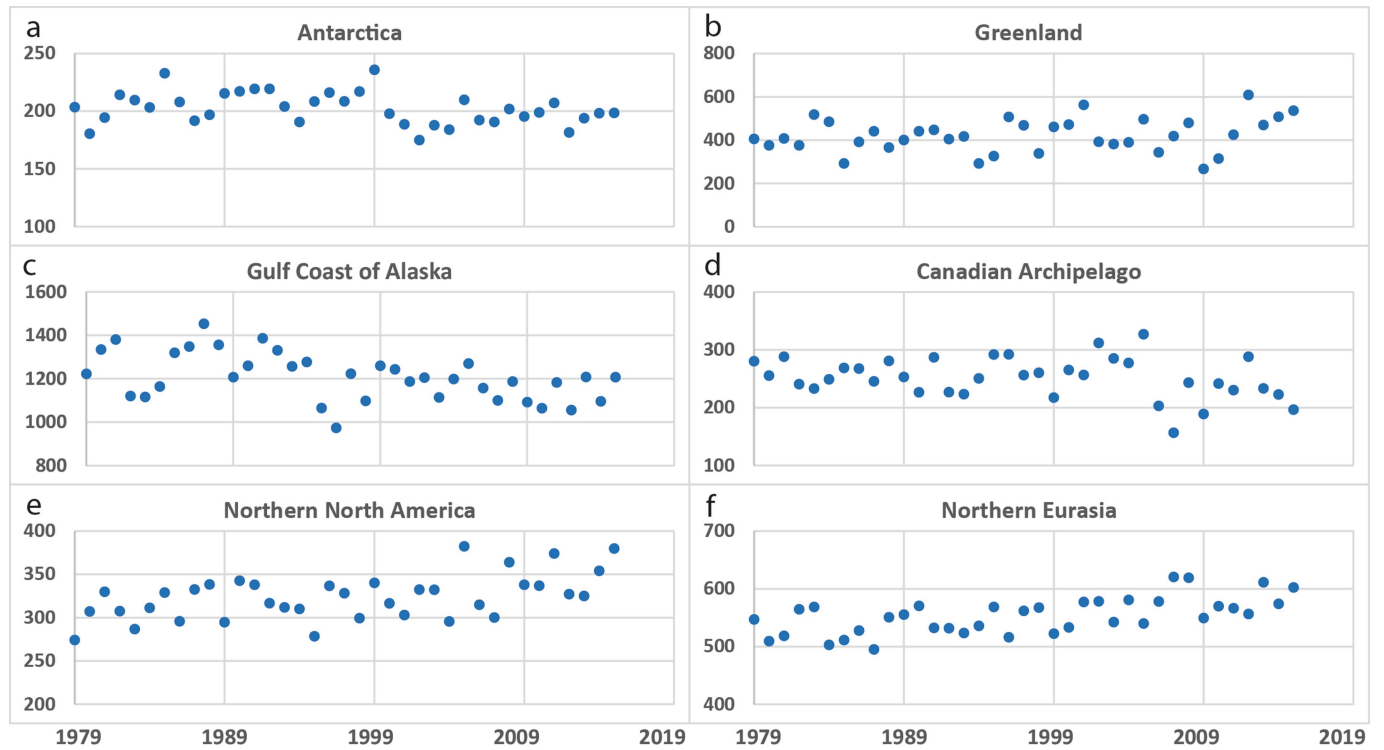
Extended Data Fig. 2 | Non-seasonal TWS anomalies—Eurasia. a–l, As in Extended Data Fig. 1, for regions 7–18.



Extended Data Fig. 3 | Non-seasonal TWS anomalies—North and South America. a–h, As in Extended Data Fig. 1, for regions 19–26.

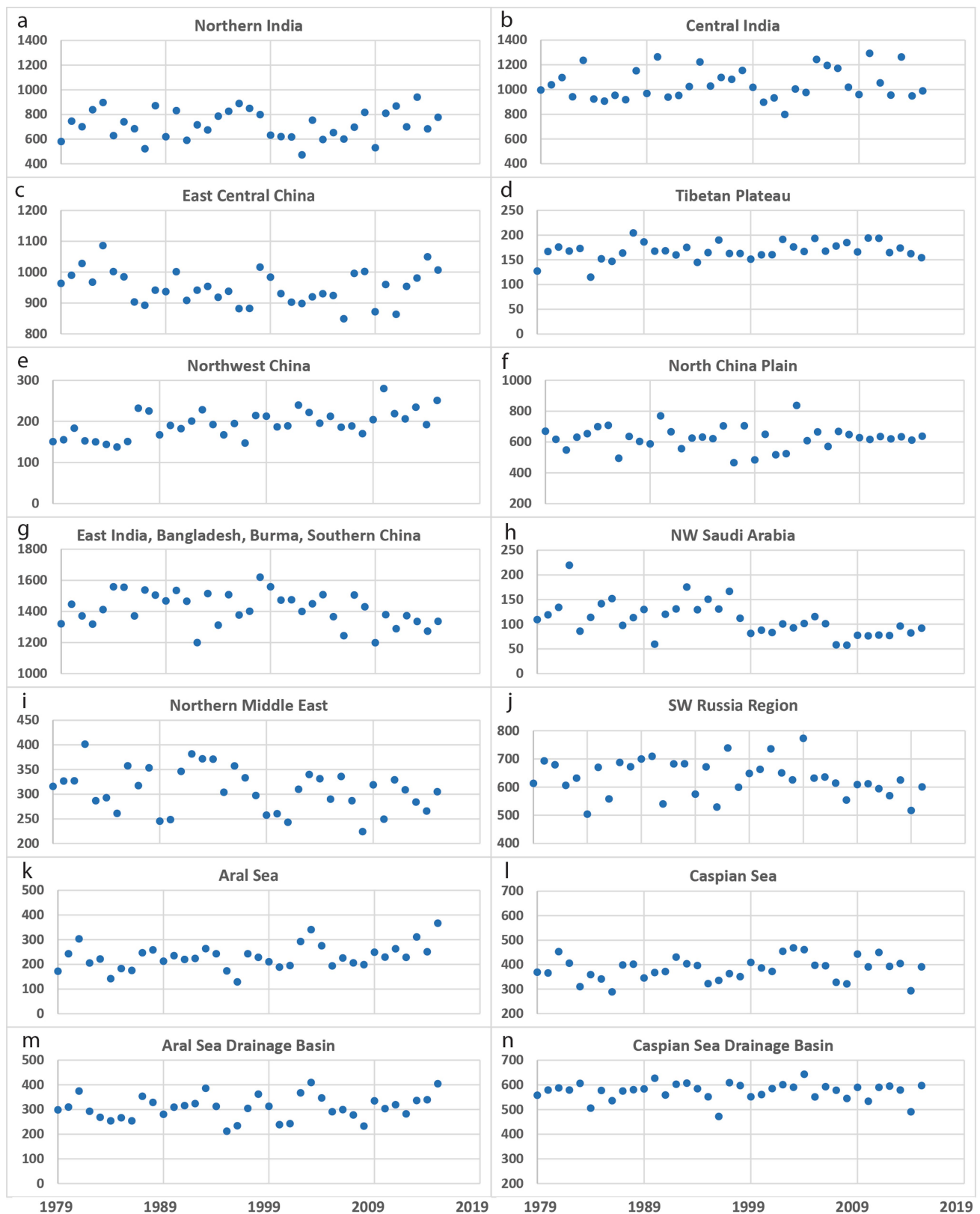


Extended Data Fig. 4 | Non-seasonal TWS anomalies—Africa and Australia. a–h, As in Extended Data Fig. 1, for regions 27–34.

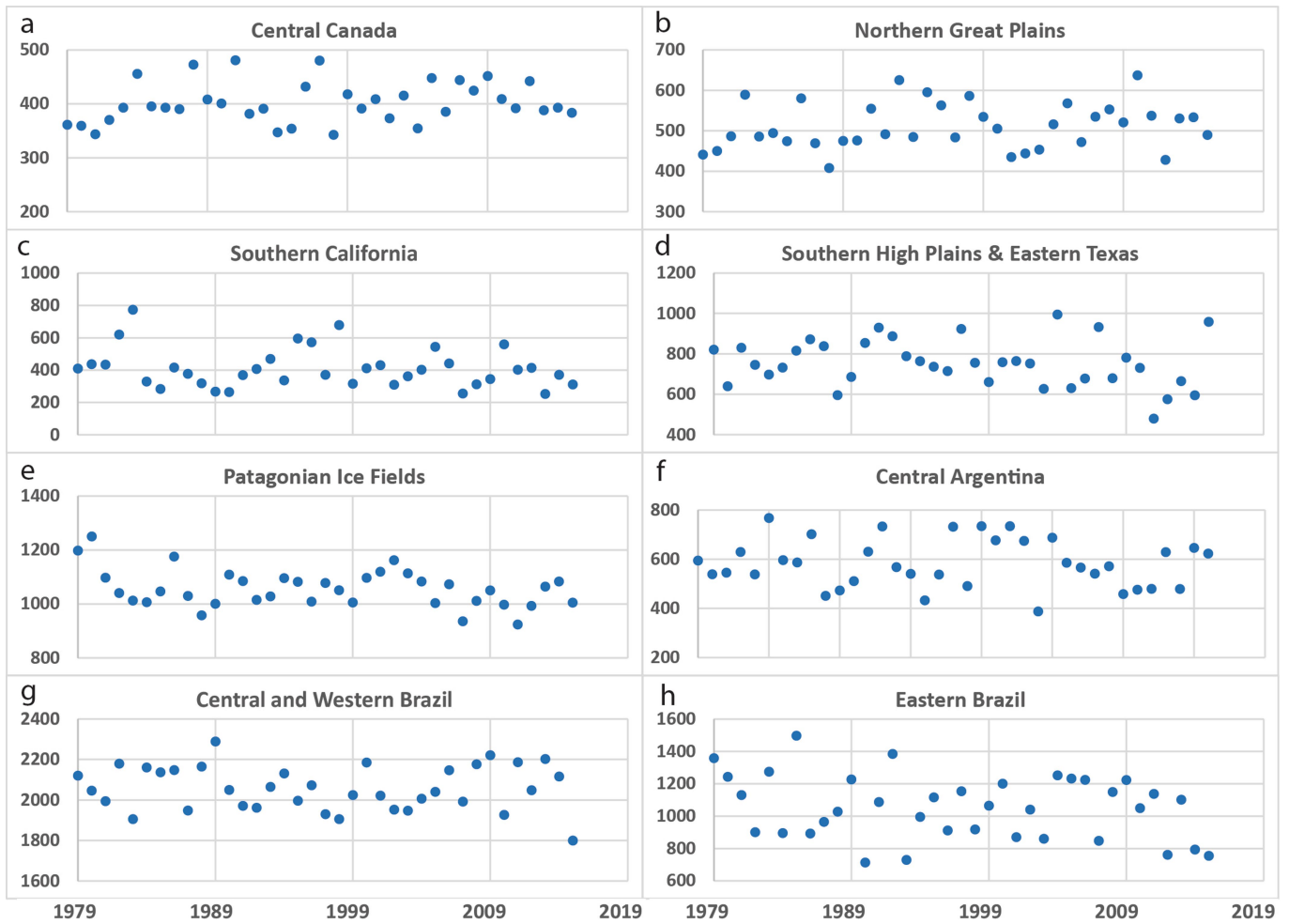


Extended Data Fig. 5 | Annual precipitation totals—global regions.
a–f, Time series of annual precipitation totals (in millimetres) averaged

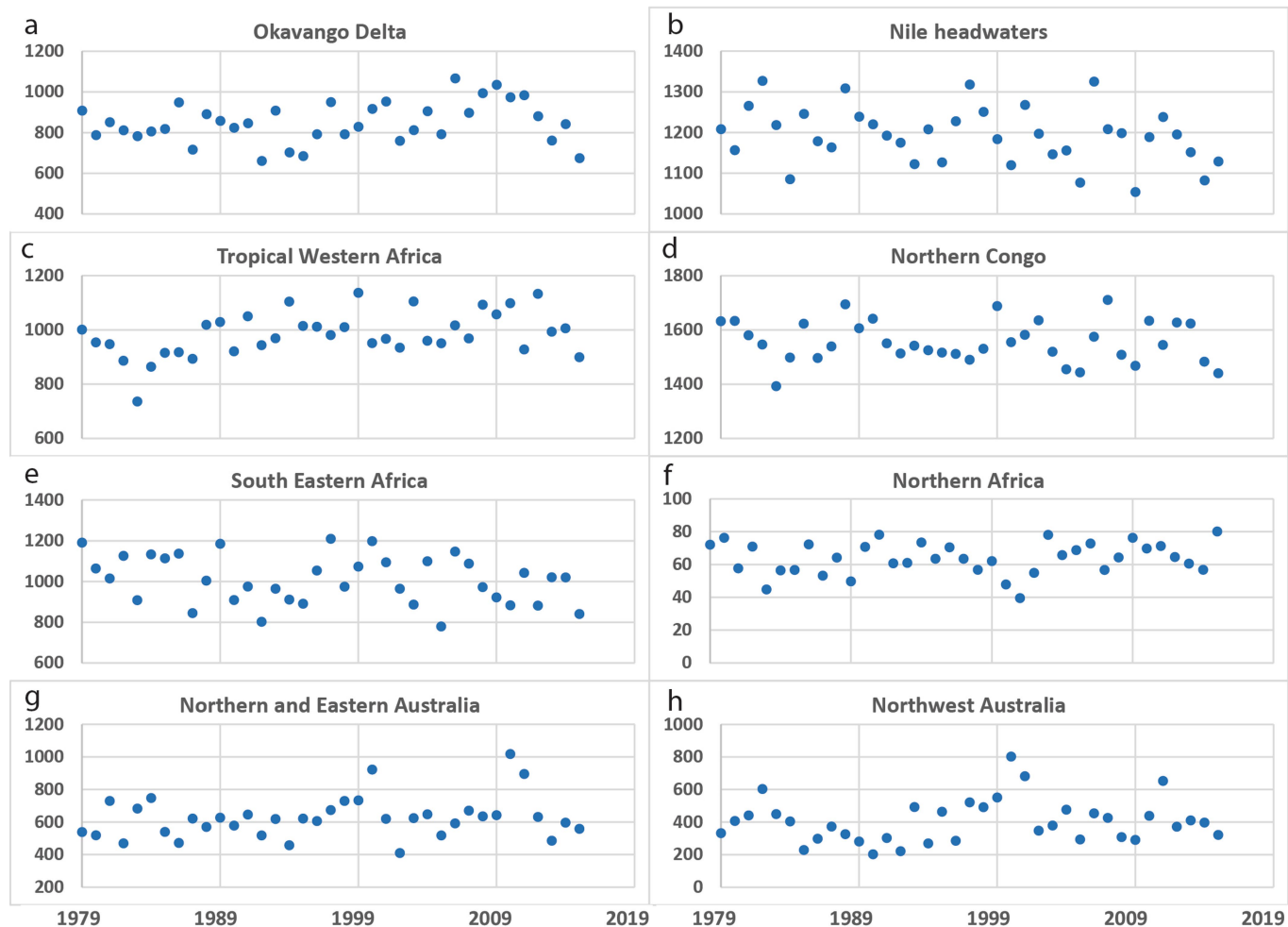
over each of study regions 1–6, based on GPCP v.2.3. We note that the y axes vary among panels.



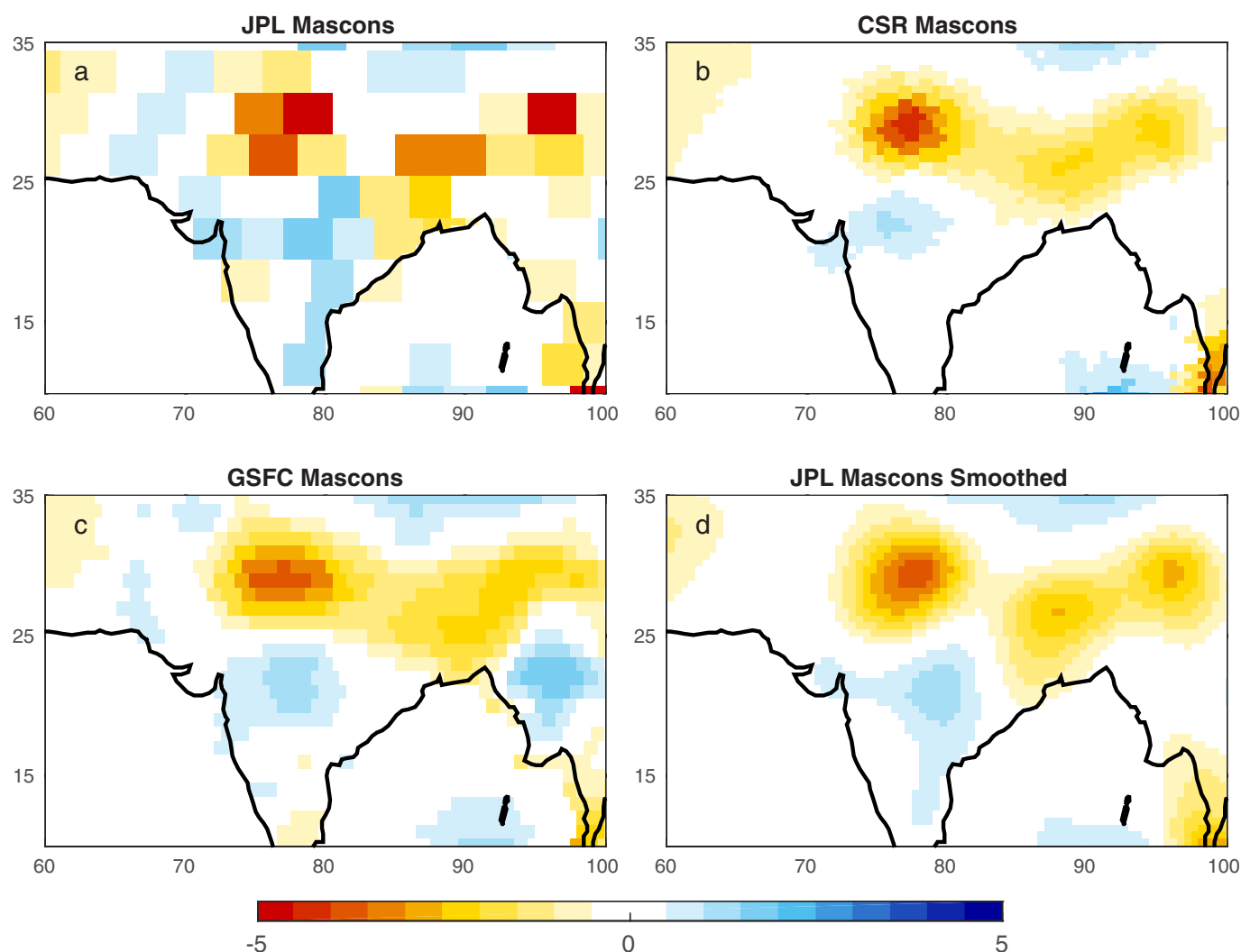
Extended Data Fig. 6 | Annual precipitation totals—Eurasia. a–n. As in Extended Data Fig. 5, for regions 7–18 and the full drainage basins of the Aral and Caspian seas.



Extended Data Fig. 7 | Annual precipitation totals—North and South America. a–h, As in Extended Data Fig. 5, for regions 19–26.

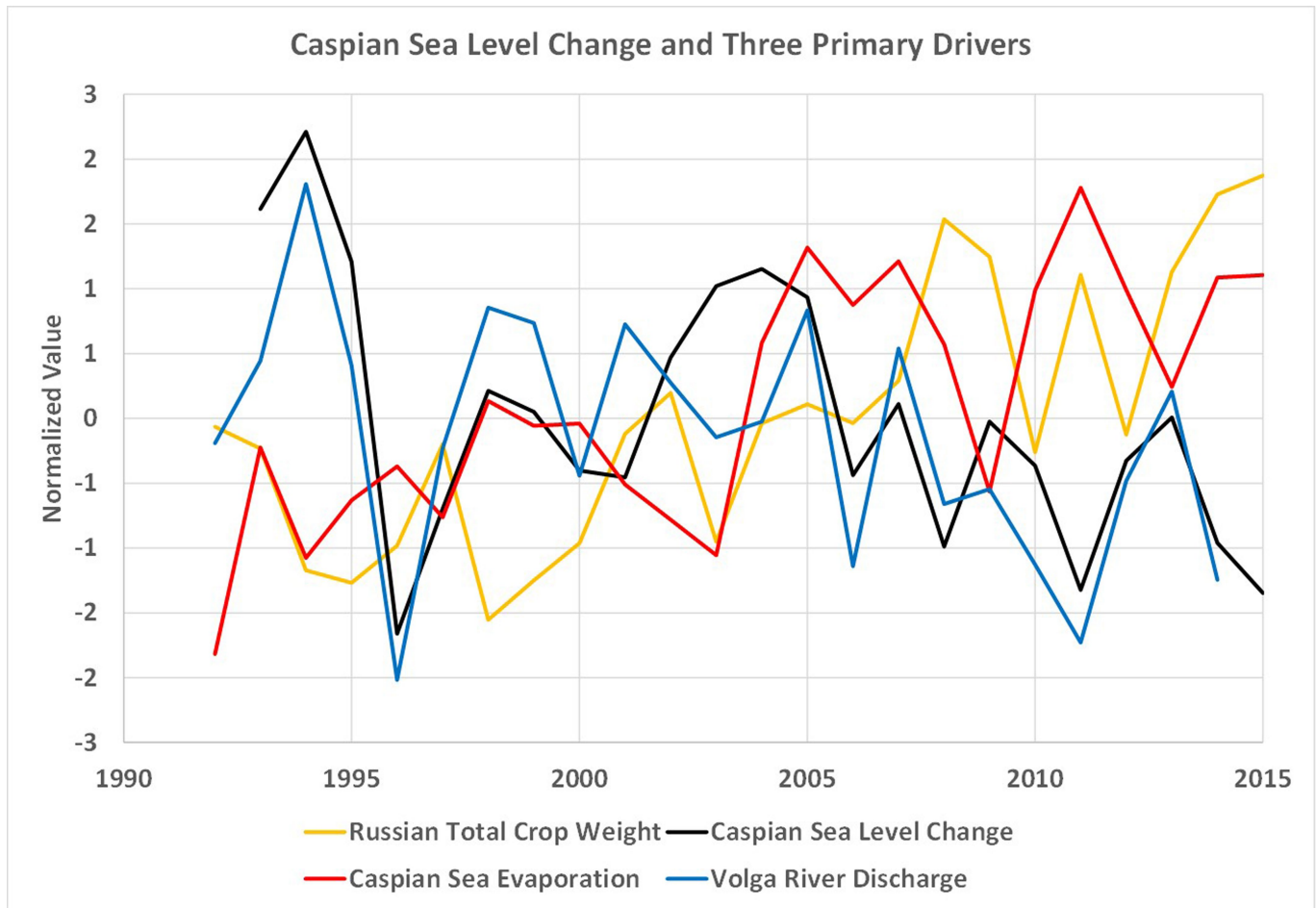


Extended Data Fig. 8 | Annual precipitation totals—Africa and Australia. a–h. As in Extended Data Fig. 5, for regions 27–34.



Extended Data Fig. 9 | Comparison of TWS trends (in centimetres per year) over India (January 2003 – March 2016) from three GRACE mascon solutions. a–d, JPL-M 3° (a), CSR-M 1° (b), GSFC-M 1° (c) and

JPL-M smoothed with a 200-km-radius Gaussian filter and plotted at 1° (d). We note the similarity between b–d, whose regional trend amplitudes have all been dampened by smoothing.



Extended Data Fig. 10 | Comparison of normalized anomalies of Caspian Sea level changes and three primary drivers. Normalized anomalies of changes in annual mean Caspian Sea level (black), Volga River discharge (blue), Russian total crop weight (yellow) and Caspian

Sea evaporation (red). Precipitation (Extended Data Fig. 6) is the other primary driver. Sea-level change is positively correlated with Volga River discharge and negatively correlated with Russian crop weight and evaporation.

Simulating the vibrational quantum dynamics of molecules using photonics

Chris Sparrow^{1,2}, Enrique Martín-López³, Nicola Maraviglia¹, Alex Neville¹, Christopher Harrold¹, Jacques Carolan⁴, Yogesh N. Joglekar⁵, Toshikazu Hashimoto⁶, Nobuyuki Matsuda⁷, Jeremy L. O'Brien¹, David P. Tew⁸ & Anthony Laing^{1*}

Advances in control techniques for vibrational quantum states in molecules present new challenges for modelling such systems, which could be amenable to quantum simulation methods. Here, by exploiting a natural mapping between vibrations in molecules and photons in waveguides, we demonstrate a reprogrammable photonic chip as a versatile simulation platform for a range of quantum dynamic behaviour in different molecules. We begin by simulating the time evolution of vibrational excitations in the harmonic approximation for several four-atom molecules, including H₂CS, SO₃, HNCN, HFHF, N₄ and P₄. We then simulate coherent and dephased energy transport in the simplest model of the peptide bond in proteins—*N*-methylacetamide—and simulate thermal relaxation and the effect of anharmonicities in H₂O. Finally, we use multi-photon statistics with a feedback control algorithm to iteratively identify quantum states that increase a particular dissociation pathway of NH₃. These methods point to powerful new simulation tools for molecular quantum dynamics and the field of femtochemistry.

Early electronic computers exploited analogies with acoustic, thermal or mechanical phenomena, such as capacitance for spring stiffness, to simulate a range of practically relevant physical systems. Whereas modern digital simulations have become versatile foundational tools in science and engineering, all classical computers are fundamentally inefficient at tackling exponentially complex microscopic behaviour such as the quantum dynamics of molecules^{1,2}. A proposed solution is to engineer quantum mechanical components into devices that are then inherently capable of simulating quantum systems^{3–6}. Here, we demonstrate how integrated quantum photonics can be used to develop simulation methods for molecular quantum dynamics, by building on the analogies between optical modes in waveguides and vibrational modes in molecules and between single photons and quantized vibrational excitations.

Advances in the control of ultrafast molecular dynamics have revealed the importance of quantum interference among vibrational modes in behaviour such as bond-selective chemistry². In applying optimal control theory to a harmonic model of chained atoms⁷, it has been shown in principle how a control field could drive the dynamics of quantum interference between vibrational modes⁸ to excite local bonds. However, laboratory demonstrations of selective bond dissociation required adaptive feedback control to put the principles into practice⁹. Further control over vibrational wavepackets has enabled selective dissociation governed by a single quantum of vibrational energy¹⁰, manipulation of individual molecules at ambient conditions¹¹, preparations of coherent superpositions on sub-femtosecond timescales¹², and single vibrational states of ultracold molecules¹³. Molecular dynamics are now observable on their ultrafast intrinsic timescale¹⁴.

The prospect of more sophisticated control with quantum states of light and for larger molecules increases the challenge of simulating dynamic behaviour. Light–matter interaction with squeezed states has been demonstrated experimentally in several contexts (see, for example, ref. ¹⁵); enhanced spectroscopy and the control of molecules with

multi-mode, multi-photon states has been shown theoretically (see, for example, ref. ¹⁶), with techniques for pulse shaping of quantum states of light also demonstrated (see, for example, ref. ¹⁷). Evolving a multi-excitation state across many vibrational modes is computationally inefficient even for the basic model in which normal modes are described as independent quantum harmonic oscillators. Owing to their bosonic nature, the probability amplitudes for input–output transitions among the modes are determined by matrix permanents, the calculation of which is generally extremely complex¹⁸. More detailed molecular models, for example, with anharmonic corrections to the potentials, are also likely to be computationally complex.

Quantum algorithms for the efficient simulation of Hamiltonian dynamics^{4,19} have been a strong motivator for digital quantum computers, such as those that use trapped ions²⁰. Promising digital algorithms for simulating reaction dynamics²¹ and obtaining thermal rate constants²² have been presented that harness the exponential quantum speed-up. Yet, achieving fault tolerance²³ and the high logical-gate counts²⁴ that enable these applications is extremely challenging. Ansatz-based methods, such as the variational approach for solving the eigenvalue problem²⁵, have reduced demands, as demonstrated recently with superconducting qubits²⁶, but the difficulties associated with applying such an approach to Hamiltonian dynamics have yet to be overcome. Analogue quantum simulations⁶, in which a quantum system of interest can be mapped directly onto a quantum technological platform, may enable practical advantages in the nearer-term.

Progress in photonic quantum technologies over the past decade has seen the introduction of on-chip processing of photonic quantum information^{27–29}, full reprogrammability for linear optical circuitry³⁰, and the integration of photon generation^{31,32} and detection³³. Solid-state single-photon sources³⁴ and high-efficiency detectors³⁵ have recently been demonstrated as a solution to achieving large numbers of photons. Ultimately, basic methods to correct for photon loss are likely to be required before photonic quantum simulations outperform

¹Quantum Engineering and Technology Laboratories, School of Physics and Department of Electrical and Electronic Engineering, University of Bristol, Bristol, UK. ²Department of Physics, Imperial College London, London, UK. ³Nokia Bell Labs, Cambridge, UK. ⁴Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. ⁵Department of Physics, Indiana University Purdue University Indianapolis (IUPUI), Indianapolis, IN, USA. ⁶NTT Device Technology Laboratories, NTT Corporation, Atsugi, Japan. ⁷NTT Basic Research Laboratories, NTT Corporation, Atsugi, Japan. ⁸School of Chemistry, University of Bristol, Bristol, UK. ⁹These authors contributed equally: Chris Sparrow, Enrique Martín-López. *e-mail: anthony.laing@bristol.ac.uk

classical algorithms³⁶, but the demands on error correction for specialized quantum simulators could be much lower than those for universal digital quantum simulators³⁷. Here, our focus is on establishing programmable linear optical circuitry as a core capability for simulating the vibrational dynamics of the atoms within molecules.

Simulation procedure

Diagonalizing the Hessian matrix of a molecule in mass-weighted coordinates provides its vibrational spectrum and normal modes, which define a Hamiltonian of independent quantum harmonic oscillators:

$$\hat{H} = \sum_i \hbar \omega_i a_i^\dagger a_i$$

where \hbar is the reduced Planck constant, ω_i is the angular frequency of the i th mode, and a_i^\dagger and a_i are the bosonic creation and annihilation operators of the i th mode. The spatial localization of vibrational energy is important for understanding many molecular phenomena, such as energy transport and dissociation. We therefore consider a basis transformation

$$a_i^\dagger \rightarrow \sum_k U_{ki}^L a_k^\dagger$$

where U^L is a unitary matrix, to a set of modes localized around a single atomic site or chemical bond. Dynamics in the localized basis can then be simulated via the model Hamiltonian

$$\hat{H}_L = \sum_{k,j} H_{kj}^L a_k^\dagger a_j$$

where

$$H_{kj}^L = \sum_i \hbar \omega_i U_{ki}^L \overline{U_{ji}^L}$$

and the overbar denotes complex conjugation.

This general model can be simulated directly for m vibrational modes of any given molecule with a linear optical chip that can be programmed to implement any unitary operation over m modes. Reconfiguring such a device to implement the unitary transfer matrix $U(t_i) = \exp(-iH^L t_i/\hbar)$ for a series of time steps $\{t_i\}$ enables simulation of the Hamiltonian \hat{H}_L on any initial multi-mode vibrational state via its mapping to a multi-mode optical input state. Here, we use a silica-on-silicon integrated photonic chip that is fully programmable over six waveguides via 30 thermo-optic phase shifters³⁰ to perform molecular simulations of up to six-mode vibrational systems. We simulate initial states of up to four vibrational quanta, with states of up to four single photons, produced from spontaneous parametric down-conversion sources. Photons are detected with single-photon counting modules. The number and pattern of photons collected at the output of the optical modes for each circuit configuration are governed by the probabilities for the molecule to be found in the corresponding vibrational states at the corresponding time step.

Vibrational dynamics of four-atom molecules

Thioformaldehyde (H_2CS), a key molecule for spectroscopic experiments, is shown in Fig. 1a with its normal-mode spectrum. The six localized vibrational modes of H_2CS comprise two CH stretch modes, two CH bend modes, a CS stretch mode and a wagging mode, which are mapped to our photonic chip from the normal-mode basis, as depicted conceptually in Fig. 1b. We initialized the simulation for the state $|\mu\rangle \propto \mu |1_{\text{CHs1}}, 1_{\text{CHs2}}\rangle + \mu^2 |2_{\text{CHs1}}, 2_{\text{CHs2}}\rangle$ (with small squeezing parameter μ), which consists of multiple excitations superposed over the two CH stretch modes ('CHs1' and 'CHs2'), by injecting the two-mode squeezed vacuum state that was produced by the spontaneous parametric downconversion source, into the two waveguides that correspond to the CH stretch modes. Photons were collected over a series of circuit configurations that correspond to time steps of the H_2CS

local-basis Hamiltonian. In Fig. 1c we display the experimentally simulated evolution of the probabilities for excitations to be found in only the CH stretch modes, in only the CH bend modes and shared between these stretch and bend modes, for the two-excitation (upper panel) and four-excitation subspace (lower panel).

Dynamics in the two-excitation subspace involve both excitations oscillating between stretch and bend modes via the intermediate state in which one excitation is in each of the subspaces. The L^1 distance

$$\mathcal{D}(\mathbf{p}, \mathbf{q}) = \frac{1}{2} \sum_i |p_i - q_i|$$

between the results for an experimentally simulated time step (\mathbf{p}) and the ideal distribution (\mathbf{q}) is averaged over all time steps to give $\bar{\mathcal{D}} = 0.06 \pm 0.03$. In the four-excitation subspace, in which both of the stretch modes are initially doubly occupied, the experimentally simulated evolutions of probabilities for both stretch modes to remain doubly occupied, for both bend modes to become doubly occupied, and for combinations of one doubly occupied stretch mode and one doubly occupied bend mode are shown. The apparent damping of the oscillatory behaviour between these probabilities is attributable to the combinatorially growing space of multiple excitations available to the evolving state. The distance between the experimentally simulated and ideal evolutions for the full four-photon distributions, averaged over all time steps, is $\bar{\mathcal{D}} = 0.16 \pm 0.07$. The full distributions for these and all subsequent experiments are provided in Supplementary Information.

Because time is a programmable parameter in our simulator, we are able to study molecular vibrations whose evolution involves different timescales, such as the local CH stretch mode in H_2CS , which is a superposition of normal modes with lower and higher frequencies. The probability for a single excitation to remain localized in a CH stretch mode was simulated on two timescales that differ by an order of magnitude. Heralded single photons were injected into the mode that corresponds to a local CH stretch. The circuit was programmed to implement sets of unitary transformations that correspond to a short (30 fs) high-resolution window and that correspond to a longer (300 fs) low-resolution window, the behaviour of which can be observed by averaging over the high-resolution windows. In Fig. 1d we display data for these simulations, which exhibit both higher- and lower-frequency oscillations. Averaging over both evolutions gives a mean distance of $\bar{\mathcal{D}} = 0.014 \pm 0.006$.

Our six-mode simulator can explore the full space of vibrational dynamics for a general molecule of up to four atoms, as we demonstrate for P_4 , SO_3 , HNCO , HFHF and N_4 . In Fig. 1e–i we show the time evolution of a single excitation initially prepared in a local stretch mode. The change in the occupation probability to a second, spectrally overlapped (coupled) local mode is plotted. We observe dynamics with varying characteristic times governed by the vibrational spectra of the molecules. Owing to its geometry and bonding structure, P_4 has the longest-period oscillations between opposing stretches, with SO_3 showing similar stretch-mode coupling behaviour on shorter timescales. By contrast, HNCO and HFHF display faster dynamics with increased mode coupling between hydrogen-bond stretches and bends. In Fig. 1i we show the dynamics of both a single excitation and two excitations initially prepared in stretch modes of N_4 . The additional structure in the vibrational spectrum and the introduction of multi-photon quantum interference results in a more complex time dependence of the detection probabilities. The average L^1 distance over all of these experiments is $\bar{\mathcal{D}} = 0.022 \pm 0.007$.

Decoherence and energy transfer in NMA

The flow of vibrational energy in molecules is a fundamental process for chemical reaction rates and functionality in biomolecules³⁸. The vibrational quantum dynamics of a molecule within an environment can be described by the interplay of coherent unitary evolution and incoherent dephasing that results from random fluctuations of the

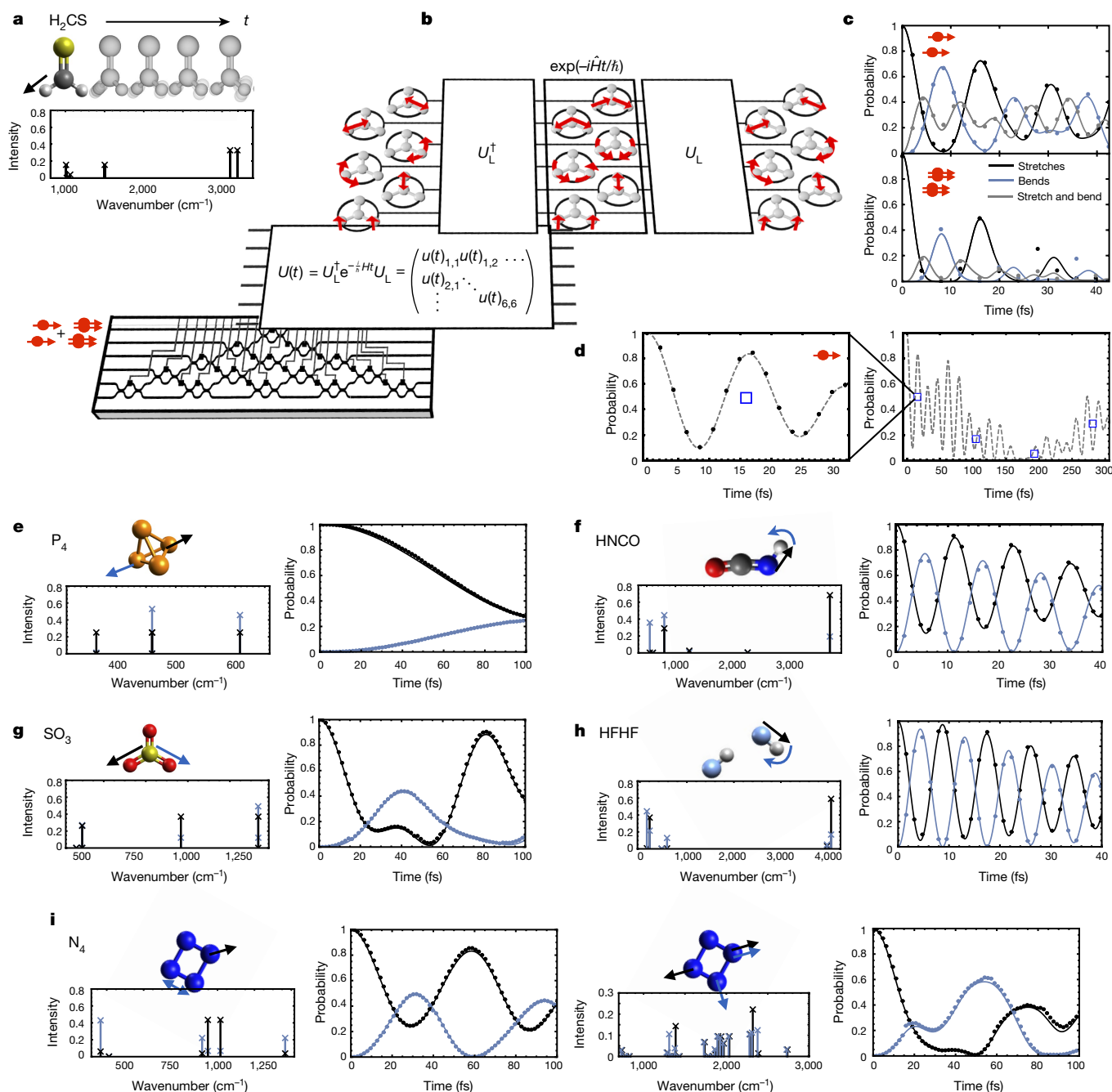


Fig. 1 | Simulating the vibrational dynamics of four-atom molecules in the harmonic approximation. **a**, Schematic evolution of a localized CH stretch mode (diagonal black arrow) in H_2CS , with its composition from normal modes plotted below. **b**, The evolution of the normal modes ($\exp(-i\hat{H}t/\hbar)$), shown schematically in the center of the top layer, is unitarily mapped (U_L and U_L^\dagger) to a set of local vibrational modes, shown schematically at the ends of the top layer. This transformation is then mapped to a time-dependent unitary transfer matrix ($U(t)$; middle layer). Simulations of photonic states under this evolution are then implemented by a reconfigurable photonic chip (bottom layer). **c**, An initial superposition of two and four excitations evolving in the localized stretch modes is simulated by injecting a two-mode squeezed vacuum state into the corresponding optical modes and collecting photon statistics for the

sequence of simulated time steps. Top and bottom panels show results for the two- and four-excitation subspaces, respectively (see insets).

d, Simulations on two timescales of the evolving probability for a single excitation to remain in a CH stretch mode. Blue squares represent the mean probability over a 30-fs window (as per left panel). **e–i**, The simulated evolution of a single excitation in P_4 (**e**), SO_3 (**g**), HNCO (**f**) and HFHF (**h**) between a local stretch mode (black) and another coupled local mode (blue). The local modes are represented diagrammatically alongside the spectral intensities of the normal modes involved. For N_4 (**i**), results are also shown for the evolution of two excitations. All data are plotted together with ideal theoretical curves; error bars displaying 1 s.d. from Poissonian statistics are very small.

vibrational frequencies—a process referred to as spectral diffusion. *N*-methylacetamide (NMA) is the simplest molecular model (Fig. 2a) of the peptide bond in proteins, where quantum coherence may have a role in energy transfer³⁹. In this section, we simulate a model for intramolecular energy transport in NMA in the presence of dephasing.

We consider a subspace that spans six backbone vibrational modes, which support a basis of approximately localized vibrational modes, including two rocking modes (curved arrows in Fig. 2) and two stretch modes (straight arrows in Fig. 2). Uniform dephasing between all modes is achieved by a time-dependent statistical averaging over the set of experiments with transfer

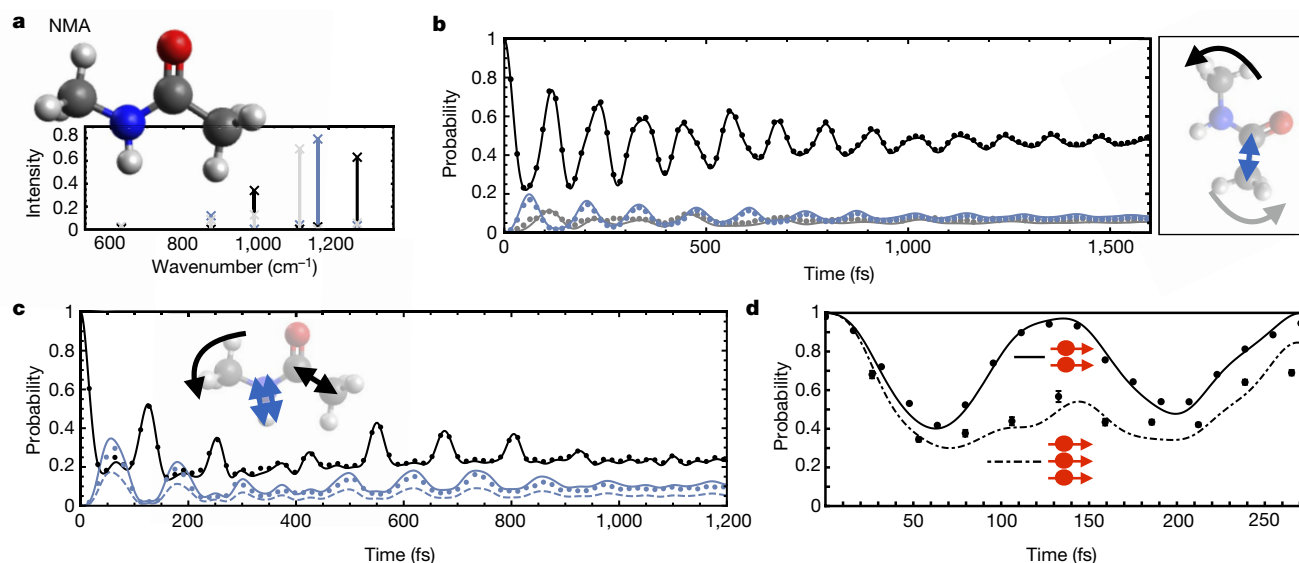


Fig. 2 | Quantum energy transfer and dephasing in NMA. **a**, A six-mode vibrational subspace of the NMA molecule is considered, with the spectral components of three localized modes colour-coded as per the arrows in **b**. **b**, Experimental simulation results for the probability of a single excitation (black points) that is initially in a local rocking mode (black arrow) at one end of the molecule and its transfer (blue and grey points) to local modes at the opposite end (blue and grey arrows) when subject to a dephasing channel with $T_2 = 0.53$ ps. **c**, Experimental simulation results for the evolution of a two-excitation state (black points) that is initially in separate local modes (black arrows) and its probability (blue points)

of being found bunched in the NH stretch mode (blue arrows) under the same dephasing channel. Solid lines represent theory. The dashed blue line plots a theoretical curve for distinguishable (or classical) excitations to be found bunched in the NH stretch mode. **d**, Experimental simulation results for the total probability of measuring a fully anti-bunched state of two excitations with the same initial state as for **c** (black points with solid black theory curve) and of measuring a fully anti-bunched state of three excitations initialized in the modes shown in **b** (black points with dot-dashed theory curve). All error bars represent 1-s.d. estimates from Poissonian statistics.

matrices $U(t_i, k) = U_L Z(k) U_L^\dagger \exp(-iH^\dagger t_i/\hbar)$, where $Z(k)$ are Heisenberg–Weyl matrices (defined in Supplementary Information) labelled by k and the average is taken over k at each time step.

Using a single photon, we simulated the probability for a single excitation initialized in a local rocking mode at one end of the molecule to be transferred to two localized modes (a rocking mode and a CC stretch mode) at the opposite end of the molecule. The experimental results shown in Fig. 2b show dynamics that are initially oscillatory, with vibrational energy transfer between the rocking modes at either end of the molecule via an intermediate CC stretch mode. The increasing effect of the suppression of coherence from dephasing results in evolution towards a steady state. Peak probabilities for energy to be localized at either end of the molecule are higher under quantum coherent dynamics than under purely ballistic classical dynamics. We used a T_2 time constant of coherence decay of 0.53 ps, but any time constant can be simulated by changing only the post-processing of data.

Simulating multiple excitations allows us to investigate the interplay of dephasing and quantum interference for multi-excitation energy transport. By injecting one photon into the waveguide that corresponds to the rocking mode and another photon into the waveguide that corresponds to the CC stretch mode, which are each localized at opposite ends of the NMA molecule (black arrows in Fig. 2c), we simulated the change in the probability for this state and for the state in which both excitations ‘bunch’ in an NH stretch mode (double blue arrows in Fig. 2c). The results in Fig. 2c show more complex oscillatory transfer between these bunched and anti-bunched states, which again tends towards a steady state. However, after full dephasing has occurred, the probability for two excitations to be bunched in the NH stretch mode is twice as high for excitations that behave as indistinguishable bosonic particles than for excitations that behave as distinguishable or classical particles (such as two excitations that pass through the molecule at different times).

For a given molecule, the probability that no bunching occurs (multiple excitations not localized around the same bond) generally decreases as the number of excitations increases⁴⁰. In Fig. 2d the probability for the subspace of no-bunching events is simulated for two and three

excitations under fully coherent dynamics. The initial state for the two-excitation evolution is the same as in the previous example; the initial state for the three-excitation evolution comprises an excitation in each of the local modes shown in Fig. 2b. The average distances across all single-, two- and three-excitation distributions in these examples are 0.017 ± 0.005 , 0.05 ± 0.01 and 0.14 ± 0.07 , respectively.

Vibrational relaxation in liquid water

We now consider extensions to the idealized model of uncoupled harmonic oscillators to account for more realistic situations, including energy dissipation and anharmonic potentials. We choose models for H₂O to demonstrate our techniques.

For a molecule that interacts with its environment, vibrational energy is exchanged via intra- and intermolecular coupling to other degrees of freedom, eventually leading to thermalization. This process is known as vibrational relaxation, and its pathways for H₂O remain an area of investigation^{41,42}. Here we simulate the relaxation of H₂O via an amplitude-damping model (Fig. 3a).

We consider a Lindblad master equation, which results in a set of time-dependent Kraus operators that can be simulated via an ensemble of transfer matrices. This evolution cannot be described as a convex sum of unitary evolutions as in the previous section; however, the transfer matrices can be realized within a unitary matrix of twice the size, via unitary dilation⁴³. Because H₂O has three vibrational modes, its three-dimensional (non-unitary) transfer matrices can be realized within a six-dimensional unitary matrix and implemented on our six-mode chip (Fig. 3b). We used experimentally measured relaxation times $\{T_i\}$ for liquid water at room temperature⁴¹ in the model.

We simulated the thermalization of an excitation in a local OH stretch mode via the symmetric bend normal mode to its ground state of no excitations. In Fig. 3c we show the probability of measuring the excitation in the two local stretch modes (left panel) and the symmetric bend mode (right panel). Oscillations between the two high-energy stretch modes decay as the population is transferred via the lower-energy bend mode to the ground state. The average L^1 distance in these experiments was $\bar{D} = 0.024 \pm 0.007$.

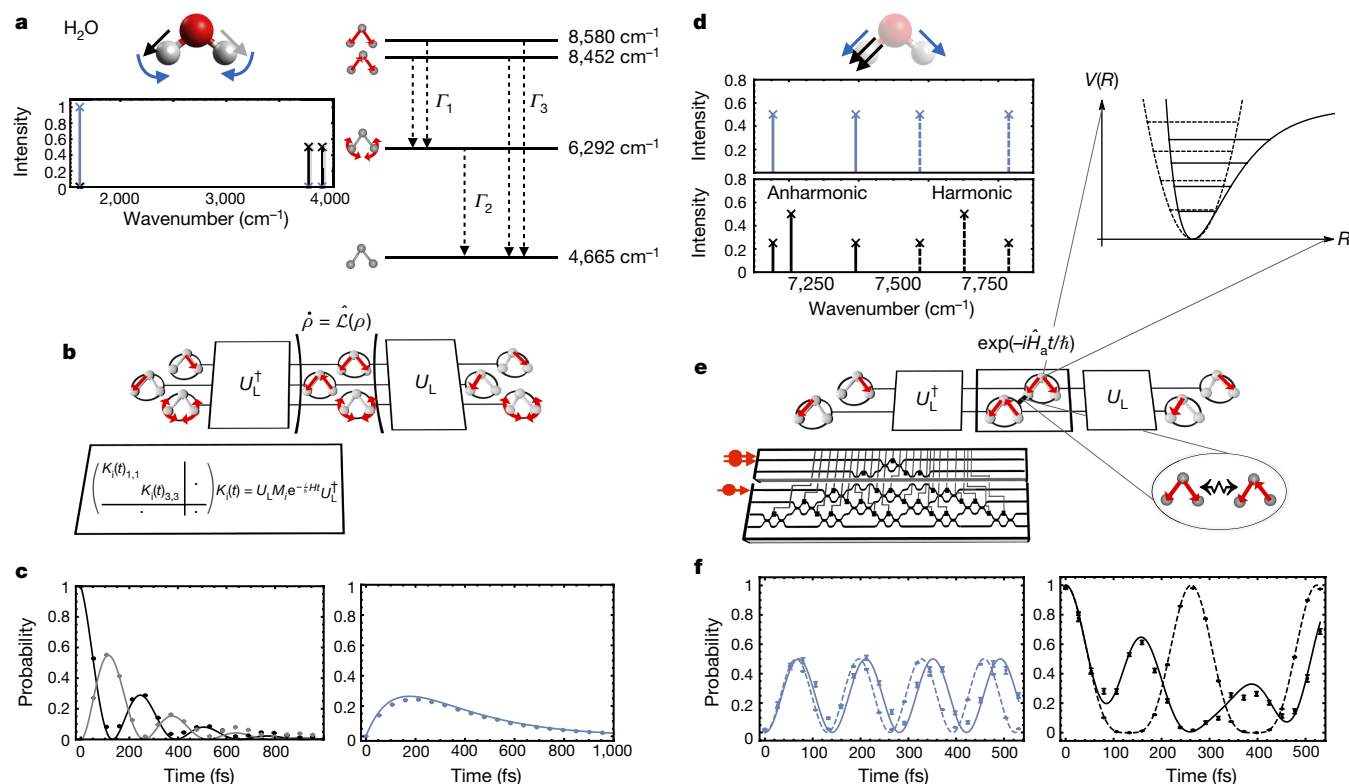


Fig. 3 | Vibrational relaxation and anharmonic evolution for H₂O.

a, Energy-level diagram for single-excitation harmonic levels and the ground state of H₂O (right) along with the spectral components of the two local OH stretch modes (black and grey) and the symmetric bend normal mode (blue) (left). $\Gamma_{1,2,3}$ are the characteristic decay rates obtained from experiments. **b**, The open-system dynamics of vibrational relaxation, described by the Lindblad equation $\dot{\rho} = \hat{\mathcal{L}}(\rho)$ where ρ is the vibrational state, can be simulated by statistically averaging the evolution under a set of linear operators implemented via unitary dilation. The Krauss operators in the localized basis, $K_i(t)$, are dilated into a unitary matrix by increasing the dimension (blocked-out parts of the matrix). **c**, Experimental results for the simulated evolution of the probability for a single excitation that is initially in one OH stretch mode (black points) to be found in the other stretch mode (grey points) and in the symmetric bend (blue points) under

Anharmonic potentials in H₂O

Potential energy surfaces of real molecules are anharmonic, and we now consider simulations in this regime, depicted in Fig. 3d. In addition to the second derivative in the Taylor expansion of the potential energy surface, as per the harmonic approximation, we now include all third derivatives and the semi-diagonal quartic derivatives. Applying vibrational perturbation theory yields a new Hamiltonian:

$$\hat{H}_a = \hat{H} + \hbar \sum_{i \leq j} \frac{x_{ij}}{2} \sqrt{\omega_i \omega_j} (a_i^\dagger a_i + a_j^\dagger a_j + 2a_i^\dagger a_j^\dagger a_i a_j)$$

where \hat{H} is the harmonic Hamiltonian and x_{ij} are the perturbation-theory coefficients.

Implementing this Hamiltonian requires interactions between photons—a key challenge in quantum information processing. Demonstrations of enhanced single-photon interactions have required, for example, an artificial Kerr medium using superconducting circuits⁴⁴, fibre coupling a single atom and a microresonator⁴⁵, or coupling to a single quantum dot in a micropillar cavity⁴⁶. Importantly, the interactions that are required to implement perturbative models such as \hat{H}_a can be weaker than the fully entangling operations and controlled π phase gates that are used for universal quantum computing, with a potentially lower demand for reprogrammable nonlinear optics.

relaxation dynamics. Solid lines are theoretical curves. **d**, Spectrum of two excitations in bunched (black) and anti-bunched (blue) local stretch modes for a harmonic (dashed) and anharmonic (solid) model. **e**, The anharmonic evolution is characterized by anharmonic potentials for single oscillators (top inset, where R is the nuclear distance and $V(R)$ is the potential energy at R) and cross-mode coupling between oscillators (bottom inset). These are implemented via a measurement-induced nonlinearity using an ancillary photon and modes. **f**, Experimental results for the simulated evolution of the probability for two excitations that are initially bunched in local stretch modes to be found in the anti-bunched state (left) and the bunched state (right) under both models (dashed, harmonic; solid, anharmonic). All error bars represent 1-s.d. estimates from Poissonian statistics.

Here, instead, we demonstrate an approach based on measurement-induced nonlinearities, which are in principle scalable for all-linear-optical quantum computing but involve a large overhead. It is possible to implement a conditional π phase shift on a two-photon Fock state using an ancillary photon and additional optical modes⁴⁷. Using newly derived nonlinear phase-shift gates, we are able to implement arbitrary phase shifts between the zero-, one- and two-photon Fock states of an optical mode.

We simulate and compare harmonic and anharmonic models of vibration for H₂O, restricting to the subspace of stretch modes. Two photons injected together into the top mode of the chip serves to simulate two excitations initialized in a superposition of the eigenstates of the harmonic model that correspond to a local OH stretch. As shown in Fig. 3e, when simulating the anharmonic model, this input state is understood as the same superposition of the new energy eigenstates of \hat{H}_a , while a third photon injected into the third optical mode serves as the ancillary system.

In Fig. 3f we show the results of simulating the probabilities for these two vibrational excitations to remain bunched or to anti-bunch, under harmonic (\hat{H}) and anharmonic (\hat{H}_a) models. The difference in the detection patterns between the two models is a function of their different spectra (Fig. 3d). The probability of detecting a single excitation in each of the modes (anti-bunched; Fig. 3f, left panel) acquires a simple frequency shift for the anharmonic evolution that corresponds to the

adjusted energy levels (Fig. 3d, top panel). By contrast, the probabilities for the state to remain doubly occupied display markedly different dynamics between the harmonic and anharmonic cases (Fig. 3f, right panel). This is a result of the three vibrational eigenstates no longer being equally spaced in energy (Fig. 3d, bottom panel), which introduces new frequencies in the evolution. For this set of experiments, the average distances between the ideal and experimental distributions for the harmonic and anharmonic cases are 0.02 ± 0.01 and 0.06 ± 0.02 , respectively.

Quantum simulation with adaptive feedback

Adaptive feedback control (AFC) is a practical laboratory technique for finding optimal control fields for molecules⁹. AFC naturally incorporates laboratory constraints to design control fields that would not be found either analytically or through numerical simulation. Nevertheless, models idealized for quantum simulation could help to identify new possibilities for molecular control, could enable their comparison over a large number of molecules and could include quantum control fields.

Our goal is to use our simulator with an adaptive feedback loop from its quantum measurement statistics to design initial quantum states for a molecule that maximally achieve a particular task over a period of evolution. Our example molecule is ammonia (NH_3), a prototype for studying dissociation, including vibrationally mediated pathways, in which the states of its products ($\text{NH}_2 + \text{H}$) depend on the prior vibrational state in the ground electronic state¹⁰.

Our model (Fig. 4b) simulates the preparation of a vibrational state in the electronic ground state of the molecule. We then obtain the vibrational state that results from an electronic excitation by projecting the vibrational modes of the ground state onto the vibrational modes of the excited state. We approximate this projection by a unitary

transformation between the modes U_{GE} ; however, this transformation can in general be achieved via single-mode squeezing, displacement and linear optical transformations⁴⁸. The evolution of the vibrational state of the electronically excited molecule is simulated under the harmonic Hamiltonian for the normal modes. By measuring the evolved state in a localized basis we identify three local NH stretch modes.

The aim of this simulation, depicted in Fig. 4c, is to let an AFC algorithm find the initial states of two vibrational excitations (in the molecule in the electronic ground state) that result in a maximal total probability of bunching in any of the three NH stretch modes (of the electronically excited molecule) over the first 10 fs of evolution, which we associate with a preferred dissociation pathway, while suppressing other bunched events, which we associate with other pathways. The algorithm begins with a trial state of two photons that simulates two excitations superposed randomly over five of the normal vibrational modes. State preparation, which is parameterized by the vector θ , is optimized iteratively by programming the simulator to implement $U(\theta, t_i) = U_L \exp(-i\hat{H}t_i/\hbar) U_{\text{GE}}^\dagger U(\theta)$, where U_{GE} relates to the transformation between the ground- and excited-state normal modes and U_L relates to the transformation between the excited-state normal and local modes.

An example experimental trial is shown in Fig. 4d. We used a Nelder-Mead simplex method to minimize the cost function

$$C = -\alpha \sum_i w_i \Delta p_i \in [-1, 1]$$

where Δp_i is the difference between the probability of bunching in the NH stretch modes and the remaining modes at time step i , w_i are weighting factors and α is a normalization factor. The final value in Fig. 4d is $C = -0.771$, starting from a random initial state with $C = +0.337$. We repeated this experiment with six random initial states;

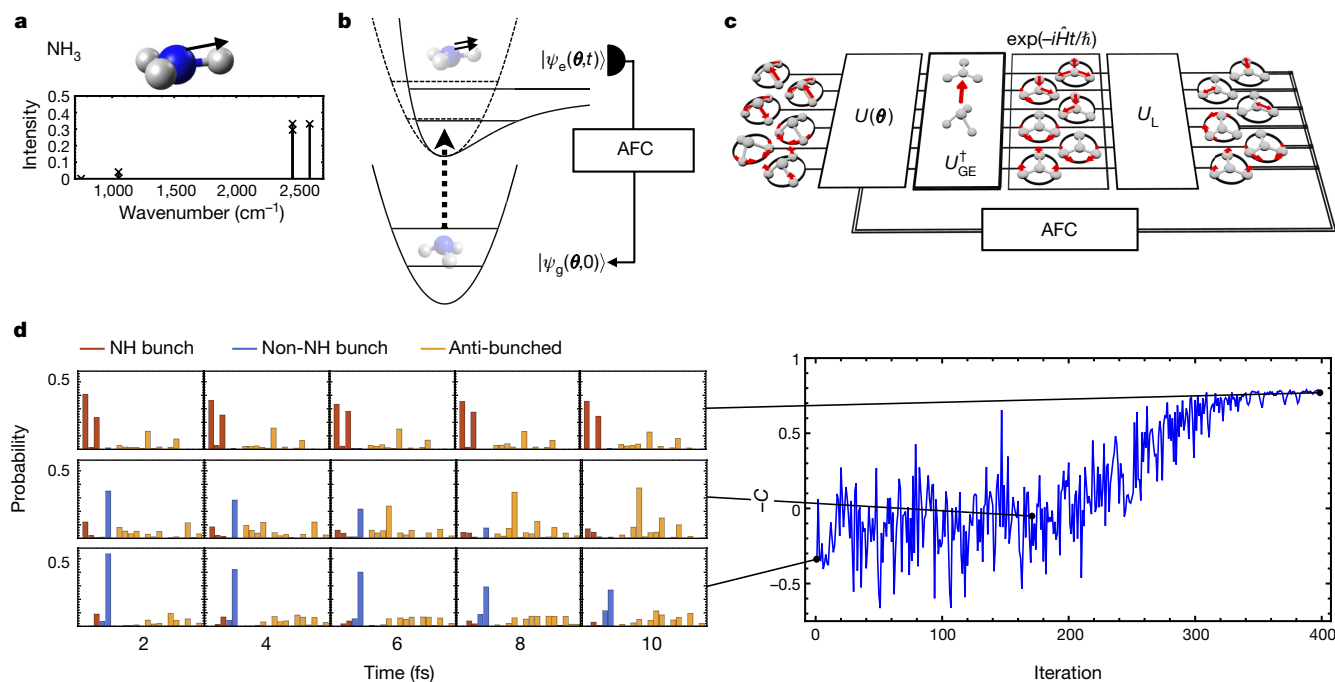


Fig. 4 | AFC algorithm for a dissociation pathway in NH_3 . **a**, The spectral decomposition of an NH stretch mode in the electronic excited state of NH_3 . **b**, A two-excitation vibrational state, parameterized by θ , is initialized in the ground-state vibrational modes ($|\psi_g(\theta, 0)\rangle$) of NH_3 . The electronic state ($|\psi_e(\theta, t)\rangle$) is excited and the localization of vibrational energy is measured over time. These measurements are used to feedback to the state preparation to increase energy localization in NH stretch modes, promoting a particular dissociation pathway for this molecule. **c**, This scenario is simulated via a parameterized unitary for state preparation $U(\theta)$, a transformation between the ground-state and excited-state modes

U_{GE}^\dagger , evolution under the excited-state modes ($\exp(-i\hat{H}t/\hbar)$) and measurement in a localized basis via U_L . The resulting photon statistics are fed back through an AFC algorithm to set θ for the next iteration (after calculating the cost function C). **d**, The left panel displays an example set of experimental results that show the full distributions for bunching in the NH stretch modes (red), bunching in the remaining three localized modes (blue) and detection in anti-bunched patterns (yellow) for five time steps at iteration numbers 1 (bottom), 175 (middle) and 399 (top). The right panel shows $-C$ measured at every iteration.

all resulted in similar final values of the cost function, with a mean of $\bar{C} = -0.845$.

Discussion

We have introduced photonics as a platform for simulating the vibrational quantum dynamics of molecules within the harmonic, perturbatively anharmonic and Linblad models, with a photonic chip playing the part of a programmable molecule. Scaling up and extending these techniques to more involved Hamiltonians with highly anharmonic atomic potentials and electronic degrees of freedom, and to realize an advantage over classical simulation techniques³⁶, presents important and interesting research directions.

Device errors that must be addressed include inevitable flaws in circuit fabrication and operation, photon distinguishability and photon loss. Although the precision that is required in the setting of any individual circuit parameter increases with dimension⁴⁹, linear optical elements with extinction of more than 60 dB have been demonstrated²⁹. Indistinguishability, or visibilities, between independent photons have been reported at 95% for on-chip sources³² and at more than 90% for time bins of a solid-state photon source³⁴. Although ultralow-loss integrated circuitry has been demonstrated⁵⁰, photon loss is a fundamental error in photonics; basic methods that alleviate some of this error would provide substantial improvements in rates for the class of experiments demonstrated here. The development of programmable nonlinear optics at the quantum level is a key functionality for quantum technologies and remains an important challenge for the field. With modest progress in these areas, our approach could yield an early class of practical quantum simulations that operate beyond current classical limits.

Data availability

The data shown in the plots and that support the findings of this study are available from the Research Data Repository of the University of Bristol at <https://doi.org/10.5523/bris.2ymwd4m50qkt26mrhpli3d1i>.

Received: 11 October 2017; Accepted: 21 March 2018;

Published online 30 May 2018.

- Gatti, F. *Molecular Quantum Dynamics*. (Springer, Berlin, 2014).
- Brif, C., Chakrabarti, R. & Rabitz, H. Control of quantum phenomena: past, present and future. *New J. Phys.* **12**, 075008 (2010).
- Feynman, R. P. Simulating physics with computers. *Int. J. Theor. Phys.* **21**, 467–488 (1982).
- Lloyd, S. Universal quantum simulators. *Science* **273**, 1073–1078 (1996).
- Aspuru-Guzik, A. & Walther, P. Photonic quantum simulators. *Nat. Phys.* **8**, 285–291 (2012).
- Georgescu, I. M., Ashhab, S. & Nori, F. Quantum simulation. *Rev. Mod. Phys.* **86**, 153–185 (2014).
- Shi, S., Woody, A. & Rabitz, H. Optimal control of selective vibrational excitation in harmonic linear chain molecules. *J. Chem. Phys.* **88**, 6870–6883 (1988).
- Shapiro, M. & Brumer, P. Coherent control of molecular dynamics. *Rep. Prog. Phys.* **66**, 859–942 (2003).
- Assion, A. et al. Control of chemical reactions by feedback-optimized phase-shaped femtosecond laser pulses. *Science* **282**, 919–922 (1998).
- Hause, M. L., Yoon, Y. H. & Crim, F. F. Vibrationally mediated photodissociation of ammonia: the influence of NH stretching vibrations on passage through conical intersections. *J. Chem. Phys.* **125**, 174309 (2006).
- Brinks, D. et al. Visualizing and controlling vibrational wave packets of single molecules. *Nature* **465**, 905–908 (2010).
- Alnaser, A. et al. Subfemtosecond steering of hydrocarbon deprotonation through superposition of vibrational modes. *Nat. Commun.* **5**, 3800 (2014).
- Tong, X., Winney, A. H. & Willitsch, S. Sympathetic cooling of molecular ions in selected rotational and vibrational states produced by threshold photoionization. *Phys. Rev. Lett.* **105**, 143001 (2010).
- Wolter, B. et al. Ultrafast electron diffraction imaging of bond breaking in di-ionized acetylene. *Science* **354**, 308–312 (2016).
- Clark, J. B., Lecocq, F., Simmonds, R. W., Aumentado, J. & Teufel, J. D. Sideband cooling beyond the quantum backaction limit with squeezed light. *Nature* **541**, 191–195 (2017).
- Dorfman, K. E., Schlawin, F. & Mukamel, S. Nonlinear optical signals and spectroscopy with quantum light. *Rev. Mod. Phys.* **88**, 045008 (2016).
- Karpiński, M., Jachura, M., Wright, L. J. & Smith, B. J. Bandwidth manipulation of quantum light by an electro-optic time lens. *Nat. Photon.* **11**, 53–57 (2017).
- Aaronson, S. & Arkhipov, A. The computational complexity of linear optics. *Theor. Comput.* **9**, 143–252 (2013).
- Berry, D. W., Childs, A. M., Cleve, R., Kothari, R. & Somma, R. D. Simulating Hamiltonian dynamics with a truncated Taylor series. *Phys. Rev. Lett.* **114**, 090502 (2015).
- Lanyon, B. P. et al. Universal digital quantum simulation with trapped ions. *Science* **334**, 57–61 (2011).
- Kassal, I., Jordan, S. P., Love, P. J., Mohseni, M. & Aspuru-Guzik, A. Polynomial-time quantum algorithm for the simulation of chemical dynamics. *Proc. Natl Acad. Sci. USA* **105**, 18681–18686 (2008).
- Lidar, D. A. & Wang, H. Calculating the thermal rate constant with exponential speedup on a quantum computer. *Phys. Rev. E* **59**, 2429–2438 (1999).
- Campbell, E. T., Terhal, B. M. & Vuillot, C. Roads towards fault-tolerant universal quantum computation. *Nature* **549**, 172–179 (2017).
- Wecker, D., Bauer, B., Clark, B. K., Hastings, M. B. & Troyer, M. Gate-count estimates for performing quantum chemistry on small quantum computers. *Phys. Rev. A* **90**, 022305 (2014).
- Peruzzo, A. et al. A variational eigenvalue solver on a photonic quantum processor. *Nat. Commun.* **5**, 4213 (2014).
- Kandala, A. et al. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature* **549**, 242–246 (2017).
- Politi, A., Cryan, M. J., Rarity, J. G., Yu, S. & O'Brien, J. L. Silicon-on-silicon waveguide quantum circuits. *Science* **320**, 646–649 (2008).
- Crespi, A. et al. Integrated photonic quantum gates for polarization qubits. *Nat. Commun.* **2**, 566 (2011).
- Harris, N. C. et al. Quantum transport simulations in a programmable nanophotonic processor. *Nat. Photon.* **11**, 447–452 (2017).
- Carolan, J. et al. Universal linear optics. *Science* **349**, 711–716 (2015).
- Silverstone, J. W. et al. On-chip quantum interference between silicon photon-pair sources. *Nat. Photon.* **8**, 104–108 (2014).
- Spring, J. B. et al. Chip-based array of near-identical, pure, heralded single-photon sources. *Optica* **4**, 90–96 (2017).
- Najafi, F. et al. On-chip detection of non-classical light by scalable integration of single-photon detectors. *Nat. Commun.* **6**, 5873 (2015).
- Wang, H. et al. Near-transform-limited single photons from an efficient solid-state quantum emitter. *Phys. Rev. Lett.* **116**, 213601 (2016).
- Marsili, F. et al. Detecting single infrared photons with 93% system efficiency. *Nat. Photon.* **7**, 210–214 (2013).
- Neville, A. et al. Classical boson sampling algorithms with superior performance to near-term experiments. *Nat. Phys.* **13**, 1153–1157 (2017).
- Cubitt, T., Montanaro, A. & Piddock, S. Universal quantum Hamiltonians. Preprint at <https://arxiv.org/abs/1701.05182> (2017).
- Leitner, D. M. Energy flow in proteins. *Annu. Rev. Phys. Chem.* **59**, 233–259 (2008).
- Kobus, M., Nguyen, P. H. & Stock, G. Coherent vibrational energy transfer along a peptide helix. *J. Chem. Phys.* **134**, 124518 (2011).
- Arkhipov, A. & Kuperberg, G. The bosonic birthday paradox. *Geometry Topology Monogr.* **18**, 1–7 (2012).
- Lindner, J. et al. Vibrational relaxation of pure liquid water. *Chem. Phys. Lett.* **421**, 329–333 (2006).
- Ramasesha, K., De Marco, L., Mandal, A. & Tokmakoff, A. Water vibrations have strongly mixed intra- and intermolecular character. *Nat. Chem.* **5**, 935–940 (2013).
- Horn, R. A. & Johnson, C. R. *Topics in Matrix Analysis* 57–59 (Cambridge Univ. Press, Cambridge, 1991).
- Kirchmair, G. et al. Observation of quantum state collapse and revival due to the single-photon Kerr effect. *Nature* **495**, 205–209 (2013).
- Shomroni, I. et al. All-optical routing of single photons by a one-atom switch controlled by a single photon. *Science* **345**, 903–906 (2014).
- De Santis, L. et al. A solid-state single-photon filter. *Nat. Nanotechnol.* **12**, 663–667 (2017).
- Knill, E., Laflamme, R. & Milburn, G. A scheme for efficient quantum computation with linear optics. *Nature* **409**, 46–52 (2001).
- Huh, J., Guerreschi, G. G., Peropadre, B., McClean, J. R. & Aspuru-Guzik, A. Boson sampling for molecular vibronic spectra. *Nat. Photon.* **9**, 615–620 (2015).
- Russell, N. J., Chakhmakhchyan, L., O'Brien, J. L. & Laing, A. Direct dialling of Haar random unitary matrices. *New J. Phys.* **19**, 033007 (2017).
- Lee, H., Chen, T., Li, J., Painter, O. & Vahala, K. J. Ultra-low-loss optical delay line on a silicon chip. *Nat. Commun.* **3**, 867 (2012).

Acknowledgements We thank A. Orr-Ewing and R. Santagati for helpful conversations, and J. Barton for assistance with figures. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC), European Commission QUCIP (H2020-FETPROACT-3-2014: quantum simulation) and the European Research Council (ERC). A.N. is grateful for support from the Wilkinson Foundation. J.C. is supported by EU H2020 Marie Skłodowska-Curie grant number 751016. Y.N.J. was supported by NSF grant number DMR-1054020. J.L.O.B. acknowledges a Royal Society Wolfson Merit Award and a Royal Academy of Engineering Chair in Emerging Technologies. Fellowship support from EPSRC is acknowledged by A.L. (EP/N003470/1).

Reviewer information Nature thanks A. Aspuru-Guzik and F. Gatti for their contribution to the peer review of this work.

Author contributions All authors contributed to discussions and project development. The concept of simulating molecular vibrations with photonics was proposed by A.L. The methodology for simulating evolutions in localized bases was developed by E.M.-L. and D.P.T., with input from A.L. and C.S. Chemical calculations were done by D.P.T., based on which C.S. and E.M.-L. developed and simulated datasets. Methods for simulating open-system dynamics were developed by C.S. with input from Y.N.J. and A.L. The concept of simulating anharmonics with nonlinear gates was proposed by A.L., with the methodology for finding the nonlinear phase shift gates developed by C.S. and A.N.; A.N. also developed this code. The concept of incorporating AFC into simulations was proposed by A.L., with methodology by C.S., N.Mar., A.N. and A.L.; A.N. also developed this code. The photonic chip was developed by N.Mat. and T.H. with input from A.L. and J.L.O'B. The experiment was built by C.H., J.C., N. Mat., N.Mar. and A.L. Data were collected by N.Mar., C.H., E.M.-L. and J.C.

Data were analysed by C.S., E.M.-L., N.Mar., A.N. and A.L. The manuscript was written by A.L., C.S. and E.M.-L. with input from D.P.T. and N.Mar. The project was conceived and managed by A.L.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0152-9>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to A.L.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Pyramidal cell regulation of interneuron survival sculpts cortical networks

Fong Kuan Wong^{1,2,3}, Kinga Bercsenyi^{1,2,3}, Varun Sreenivasan^{1,2}, Adrián Portalés^{1,2}, Marian Fernández-Otero^{1,2} & Oscar Marín^{1,2*}

Complex neuronal circuitries such as those found in the mammalian cerebral cortex have evolved as balanced networks of excitatory and inhibitory neurons. Although the establishment of appropriate numbers of these cells is essential for brain function and behaviour, our understanding of this fundamental process is limited. Here we show that the survival of interneurons in mice depends on the activity of pyramidal cells in a critical window of postnatal development, during which excitatory synaptic input to individual interneurons predicts their survival or death. Pyramidal cells regulate interneuron survival through the negative modulation of PTEN signalling, which effectively drives interneuron cell death during this period. Our findings indicate that activity-dependent mechanisms dynamically adjust the number of inhibitory cells in nascent local cortical circuits, ultimately establishing the appropriate proportions of excitatory and inhibitory neurons in the cerebral cortex.

In the adult neocortex, approximately one-sixth of neurons are inhibitory γ -aminobutyric acid-containing (GABAergic) interneurons^{1,2}, and this ratio is relatively stable across cortical regions and species regardless of the total number of neurons^{3–6}. The cellular balance between excitation and inhibition is critical for brain function and is likely to be disrupted in a number of neuropsychiatric conditions^{7–9}. However, the mechanisms that regulate the establishment of appropriate numbers of excitatory and inhibitory neurons in the cerebral cortex remain largely unknown.

Programmed cell death, also known as apoptosis, is an essential mechanism that sculpts the central and peripheral nervous systems during development^{10–12}. The death of developing neurons is mediated by an evolutionarily conserved signalling pathway that involves the pro-apoptotic BCL2 family members BAX and BAK¹³. Previous studies have shown that both cortical pyramidal cells and GABAergic interneurons undergo extensive cell death during postnatal development^{14,15}, which suggests that apoptosis may contribute to the establishment of balanced networks of excitatory and inhibitory neurons in the cerebral cortex. However, the temporal relationship and interdependency of the programmed cell death periods for both populations of neurons have not been explored in detail.

Concatenated waves of neuronal death

To determine the developmental sequence that establishes the final ratio of excitatory and inhibitory neurons in the cerebral cortex, we estimated the absolute numbers and relative proportions of pyramidal cells and GABAergic interneurons at different postnatal stages of development using stereological methods in mouse strains in which specific classes of neurons are irreversibly labelled. We chose this method to estimate programmed cell death over the direct quantification of dying cells because classical apoptotic markers such as cleaved caspase-3 have non-apoptotic roles in neurons¹⁶ and are expressed only transiently (Extended Data Fig. 1a, b). We crossed *Nex^{Cre/+}* and *Nkx2-1-Cre* mice with appropriate reporter strains (see Methods) to identify pyramidal cells (*Nex^{Cre/+};Fucci2*) and GABAergic interneurons (*Nkx2-1-Cre;RCL^{tdT}*), respectively. Expression of Cre under the control of the *Neurod6* locus in *Nex^{Cre/+}* mice labels all cortical excitatory neurons

with the exception of Cajal–Retzius cells¹⁷. In *Nkx2-1-Cre* mice, Cre specifically labels interneurons derived from the medial ganglionic eminence (MGE) and preoptic area (POA), including the two largest groups of cortical GABAergic interneurons, parvalbumin (PV) and somatostatin (SST) expressing cells¹⁸.

We observed that the total number of excitatory neurons in the neocortex decreases (by about 12%) between postnatal day (P)2 and P5, and then remains stable into adulthood (Fig. 1a, b, e). The reduction in excitatory neurons affects all layers of the neocortex and not only subplate cells (Extended Data Fig. 1c–e), which are known to undergo programmed cell death during this period¹⁹. By contrast, we found that the number of interneurons is steady until P5, drops extensively between P5 and P10 (by about 30%), and then remains constant into adulthood (Fig. 1c–e). Interneuron cell loss follows the normal maturation sequence of MGE and POA interneurons²⁰, with deep layer interneurons adjusting their numbers ahead of superficial layer interneurons (Fig. 1f). These results revealed that consecutive waves of programmed cell death adjust the final ratio of excitatory and inhibitory neurons in the developing cerebral cortex.

Interneuron activity predicts cell death

Our results indicated that the adjustment of interneuron numbers is preceded by a wave of pyramidal cell death, which suggest that these two processes might be directly linked. As previous work has shown that neuronal activity and apoptosis rates are inversely correlated in the developing brain^{21–23}, we hypothesized that pyramidal cells may impact interneuron survival by increasing the activity of the cells to which they connect. We tested this idea by monitoring the activity of MGE and POA interneurons in the superficial layers of the barrel cortex (S1BF) during the period of interneuron cell death. To this end, we generated mice expressing the fluorescent reporter tdTomato and the genetically encoded calcium sensor GCaMP6s in MGE and POA interneurons (*Nkx2-1-Cre;RCL^{tdT}/GCaMP6s* mice)²⁴ and performed long-term Ca^{2+} imaging in the same interneurons from layer 2/3 in S1BF of awake, head-restrained pups (Fig. 2a). To select the most appropriate time for these experiments, we estimated interneuron cell death in S1BF during postnatal development and found comparable dynamics to the rest of

¹Centre for Developmental Neurobiology, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK. ²Medical Research Council Centre for Neurodevelopmental Disorders, King's College London, London, UK. ³These authors contributed equally: Fong Kuan Wong, Kinga Bercsenyi. *e-mail: oscar.marin@kcl.ac.uk

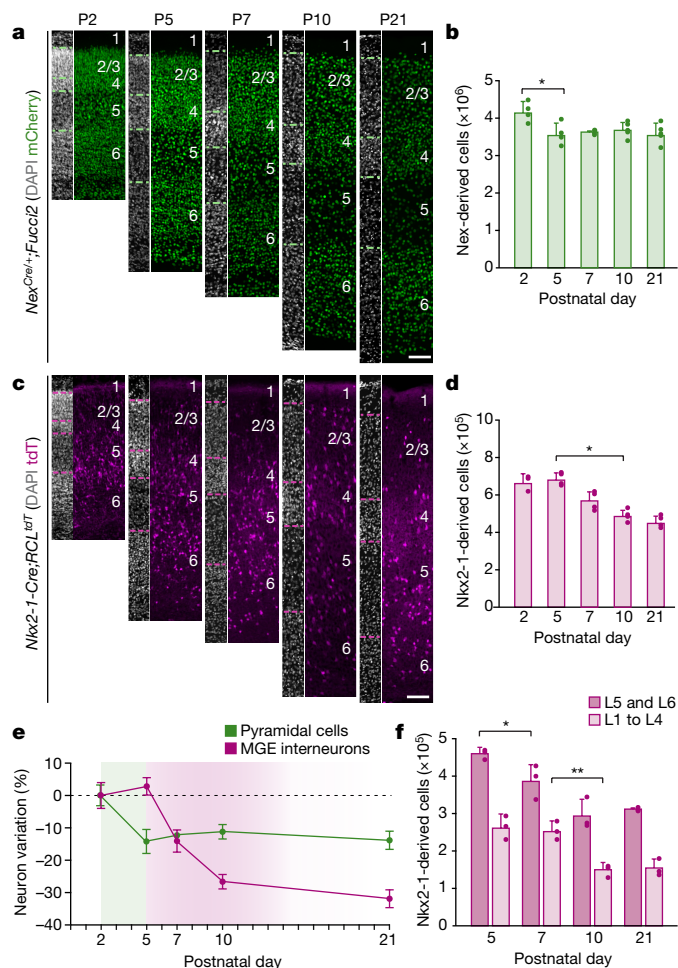


Fig. 1 | Consecutive waves of programmed cell death of pyramidal cells and interneurons in the early postnatal cortex. **a, c,** Coronal sections through the primary somatosensory cortex (S1) of *Nex^{Cre/+};Fucci2* (**a**) and *Nkx2-1-Cre;RCL^{tdTomato}* (**c**) mice during postnatal development. **b,** Total number of pyramidal cells in the entire neocortex of *Nex^{Cre/+};Fucci2* mice (ANOVA, $F = 4.17$, $*P = 0.02$; $n = 4$ (P2 and P5), 3 (P7) and 5 (P10 and P21) mice). **d,** Total number of MGE and POA interneurons in the entire neocortex of *Nkx2-1-Cre;RCL^{tdTomato}* mice (ANOVA, $F = 26.80$, $*P = 0.01$; $n = 4$ mice per age). **e,** Temporal percentage variation in pyramidal cells and MGE and POA interneurons. **f,** Total number of MGE and POA interneurons in superficial (L1–L4) and deep layers (L5 and L6) of the neocortex (two-way ANOVA, $F_{\text{interaction}} = 1.01$, $*P = 0.03$ and $**P = 0.002$; $n = 3$ animals per age). Data are shown as mean \pm s.e.m. Scale bars, 100 μm .

the neocortex (Extended Data Fig. 2). For layer 2/3, we observed the most prominent decrease in the number of MGE and POA interneurons between P7 and P8 (Fig. 2b).

We first established our ability to identify surviving interneurons at both times. As expected, the majority of *tdTomato*⁺ interneurons in a region of interest (ROI) were present in the same location the following day (Fig. 2c). However, we also observed that a fraction of *tdTomato*⁺ interneurons disappeared between P7 and P8 (Fig. 2c). As MGE and POA interneurons have ceased migration by the end of the first postnatal week²⁵, these observations are consistent with the idea that the cells that disappeared between P7 and P8 had undergone apoptosis.

We next investigated whether neural activity at P7 in interneurons that die by P8 was different from activity in cells that lived past P8. Analysis of calcium event rates (events per min) at P7 indicated that interneurons that died at P8 exhibited significantly fewer calcium events than neurons that lived past P8 (Fig. 2d, e). We next analysed whether P7 event rates could discriminate between neurons that died at P8 and neurons that lived beyond this day. Receiver-operating

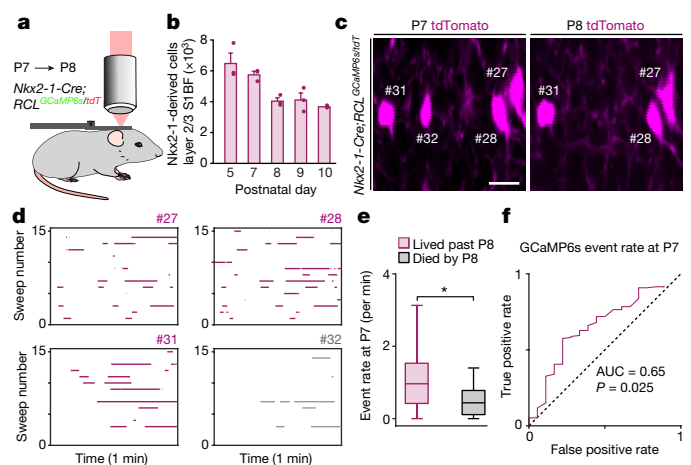


Fig. 2 | Interneuron activity levels predict cell death. **a,** Schematic of experimental design. **b,** Total number of MGE and POA interneurons in layer 2/3 S1BF of *Nkx2-1-Cre;RCL^{tdTomato}* mice ($n = 3$ animals per age). Data are shown as mean \pm s.e.m. **c,** ROI imaged at P7 (left) and P8 (right); individual neurons numbered. **d,** Raster plots showing the occurrence of calcium events at P7 for the four numbered neurons in **c**. **e,** Box plots illustrating event rates for P7 interneurons that lived past P8 (magenta) and interneurons that died by P8. In box plots, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. Two-sided Mann–Whitney test, $P = 0.03$; $n = 18$ for cells that died at P8 and 153 for cells that lived beyond P8, from three pups. **f,** ROC analysis showing the ability of P7 event rates to discriminate between cells that died by P8 and cells that lived past P8. AUC (area under the curve) = 0.65, $*P = 0.025$. Scale bar, 15 μm .

characteristic (ROC) analysis revealed that the event rate at P7 performed significantly better than chance at discriminating between these two populations (Fig. 2f). These results suggested that interneurons with relatively low levels of activity immediately before the period of interneuron cell death have an increased probability of undergoing apoptosis^{26,27}.

Pyramidal cells regulate interneuron death

The previous experiments led us to hypothesize that interneurons receiving abundant or particularly strong inputs during the period of interneuron cell death would have increased chances of survival. As PV⁺ and SST⁺ interneurons receive most of their inputs from local pyramidal cells during the first postnatal week²⁸, we reasoned that modification of the activity of cortical excitatory neurons during the period of interneuron cell death would influence interneuron survival. To test this idea, we transiently modified the activity of pyramidal cells using a chemogenetic approach based on Designer Receptors Exclusively Activated by Designer Drugs (DREADDs) that induce neuronal activation or inhibition²⁹. We injected the primary somatosensory cortex (S1) of P0 *Nex^{Cre/+}* (pyramidal cell-specific) mice with an adeno-associated virus (AAV) encoding mutant G-protein-coupled receptors that induce neuronal activation (hM3Dq) or inhibition (hM4Di) following administration of the pharmacologically inert molecule clozapine-*N*-oxide (CNO) (Fig. 3a). We then injected pups with CNO twice daily between P5 and P8, and examined the distribution of interneurons at P21 (Fig. 3a). We found that the increase in pyramidal cell activity during the period of interneuron cell death prevented this process and led to a significant increase in the density of PV⁺ and SST⁺ interneurons at P21 compared to control mice (Fig. 3b–d). This effect was not due to activity-dependent changes in the expression of PV or SST or in the density of pyramidal cells (Extended Data Fig. 3a, b). We also found that dampening the activity of pyramidal cells decreased the density of PV⁺ and SST⁺ interneurons at P21 compared to control mice, which indicates that interneuron cell death can be bidirectionally modulated by modifying the activity of pyramidal cells (Fig. 3b–d). In both experiments, changes in the density of interneurons were

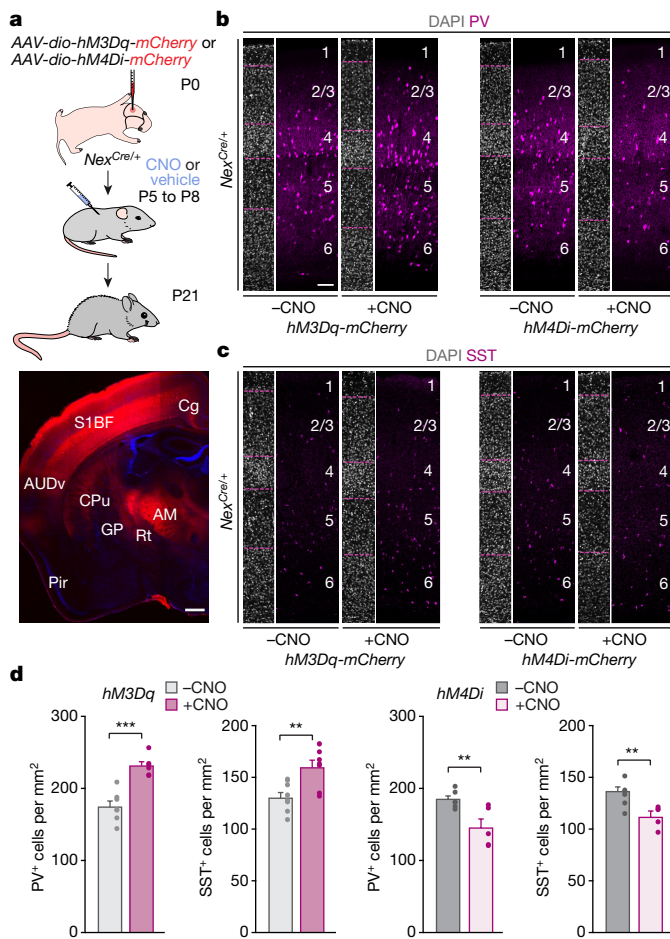


Fig. 3 | Bidirectional modulation of pyramidal cell activity regulates the extent of interneuron cell death. **a**, Top, schematic of experimental design. Bottom, mCherry expression at P21 following AAV injection at P0. AM, anteromedial thalamic nucleus; AUDv, ventral auditory cortex; Cg, cingulate cortex; CPU, caudate-putamen; GP, globus pallidus; Pir, piriform cortex; Rt, reticular nucleus. **b**, **c**, Coronal sections through S1BF from P21 *Nex^{Cre/+}* mice injected with *hM3Dq-mCherry* or *hM4Di-mCherry* viruses followed by vehicle or CNO treatment. **d**, Quantification of the density of PV⁺ and SST⁺ cells at P21. Two-tailed Student's unpaired *t*-test: for *hM3Dq*, ****P* = 0.0002, ***P* = 0.003; for *hM4Di*, ****P* = 0.006 (PV⁺), ***P* = 0.004 (SST⁺); for *hM3Dq*, *n* = 7 and 9 mice for -CNO PV⁺ and SST⁺, respectively; 6 and 7 mice for +CNO PV⁺ and SST⁺, respectively; for *hM4Di*, *n* = 7 mice for -CNO and 5 mice for +CNO for both PV⁺ and SST⁺. Data are shown as mean ± s.e.m. Scale bars, 500 μm (**a**) and 100 μm (**b**).

homogeneously distributed across layers containing PV⁺ and SST⁺ interneurons (Extended Data Fig. 3c, d). CNO administration did not cause a redistribution of interneurons from neocortical areas adjacent to the injection site (Extended Data Fig. 3e, f). Instead, we observed a prominent increase in the density of cleaved caspase-3-positive cells following inhibition of the activity of pyramidal cells during the normal period of interneuron cell death (Extended Data Fig. 4a–c). Notably, control experiments revealed that CNO did not modify the density of PV⁺ or SST⁺ interneurons in pups not infected with AAV-expressing DREADDs (Extended Data Fig. 4d, e). Similarly, CNO administration between P10 and P13 in mice injected with *hM3Dq* or *hM4Di* induced no significant changes in the density of PV⁺ and SST⁺ interneurons at P21 (Extended Data Fig. 5). Together, these results demonstrate that pyramidal cell activity is an essential regulator of interneuron survival during the normal period of interneuron cell death.

Interneurons match pyramidal cell numbers

The previous experiments suggest that pyramidal cells 'rescue' appropriate numbers of cortical interneurons from programmed cell death

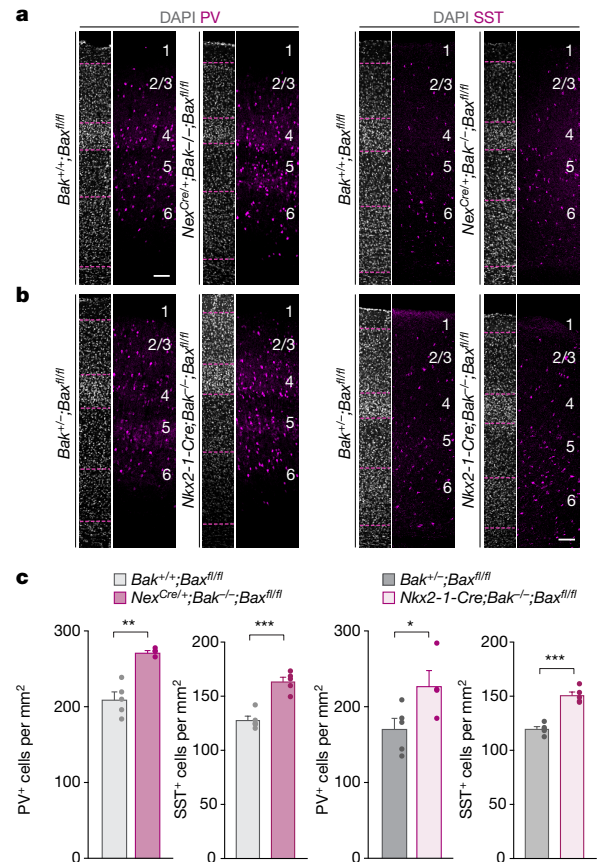


Fig. 4 | Survival of pyramidal cells rescues interneuron cell death. **a**, **b**, Coronal sections through S1BF from P30 *Bak^{+/+};Bax^{fl/fl}* and *Nex^{Cre/+};Bak^{-/-};Bax^{fl/fl}* mice (**a**), and *Bak^{+/+};Bax^{fl/fl}* and *Nkx2-1-Cre;Bak^{-/-};Bax^{fl/fl}* mice (**b**). **c**, Quantification of the density of PV⁺ and SST⁺ cells in pyramidal cell-specific *Bax/Bak* mutant mice, MGE and POA interneuron-specific *Bax/Bak* mutant mice and their respective controls at P30. Two-tailed Student's unpaired *t*-test: for *Nex^{Cre/+}*, ****P* = 0.001, ****P* = 0.0002; for *Nkx2-1-Cre*, **P* = 0.04, ****P* = 0.00004; *n* = 4 mice for *Nex^{Cre/+};Bak^{-/-};Bax^{fl/fl}* (PV⁺) and 5 mice for all other groups. Data are shown as mean ± s.e.m. Scale bar, 100 μm.

through an activity-dependent mechanism. Drawing on this idea, we reasoned that modification of the number of pyramidal cells before the period of interneuron cell death should also influence the number of surviving interneurons. To test this hypothesis, we generated conditional mice in which pyramidal cells specifically lack *Bak* and *Bax*, whose combined function is critical for apoptosis³⁰. As expected, we observed that the number of excitatory neurons in the cerebral cortex of *Nex^{Cre/+};Bak^{-/-};Bax^{fl/fl}* mutant mice did not decline between P2 and P21 (Extended Data Fig. 6). Consequently, *Nex^{Cre/+};Bak^{-/-};Bax^{fl/fl}* mutant mice had approximately 12% more pyramidal cells than control mice (Fig. 1b and Extended Data Fig. 6).

We next quantified PV⁺ and SST⁺ interneurons in S1 of control and *Nex^{Cre/+};Bak^{-/-};Bax^{fl/fl}* mutant mice at P21. The density of both PV⁺ and SST⁺ interneurons was roughly 30% higher in *Nex^{Cre/+};Bak^{-/-};Bax^{fl/fl}* mutant mice than in controls (Fig. 4a, c), which suggests that interneuron cell death is suppressed when pyramidal cell death is prevented. This increase was homogeneously distributed across layers containing PV⁺ and SST⁺ interneurons (Extended Data Fig. 7a), and was also observed in other neocortical areas (Extended Data Fig. 7c, d). To evaluate whether the increase in the number of PV⁺ and SST⁺ interneurons represented the entire population of cells that should normally have died through programmed cell death, we generated conditional mice lacking *Bax* and *Bak* in MGE and POA interneurons. We found that the density of PV⁺ and SST⁺ interneurons was also approximately 30% higher in *Nkx2-1-Cre;Bak^{-/-};Bax^{fl/fl}* mutant mice

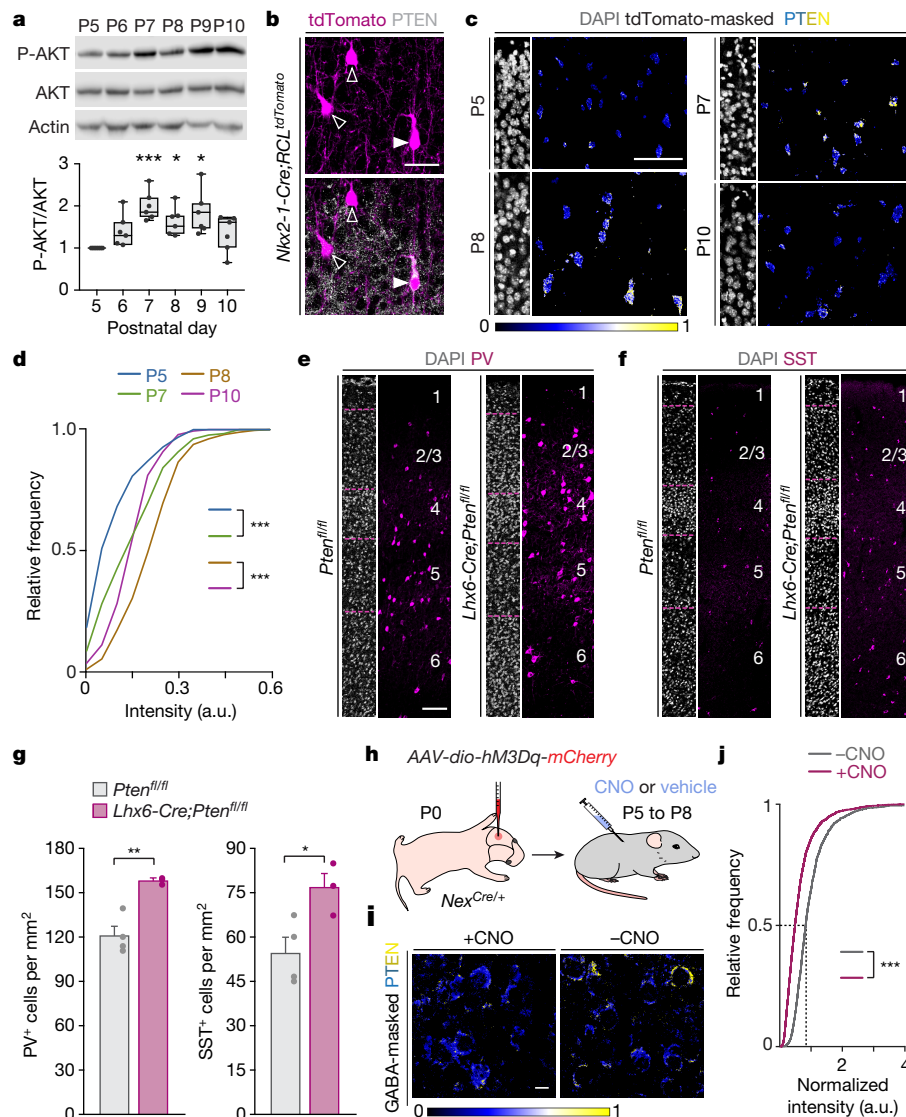


Fig. 5 | Pyramidal cell activity controls interneuron cell survival through PTEN inhibition. **a**, P-AKT, AKT and actin protein levels in the neocortex. Friedman test, $P = 0.001$; $*P = 0.03$ for P5 versus P8, $*P = 0.0101$ for P5 versus P9 and $***P = 0$ for P5 versus P7; $n = 6$ mice for each age. **b**, Coronal section through layer 2/3 S1BF from *Nkx2-1-Cre;RCL^{tdTomato}* mouse at P8. Some interneurons have much higher PTEN levels (filled arrowheads) than most (open arrowheads). **c**, Coronal sections through layer 2/3 S1BF from *Nkx2-1-Cre;RCL^{tdTomato}* mice at P5, P7, P8 and P10. PTEN expression is shown as a custom look up table (LUT, blue–white–yellow) in tdTomato-masked cells. **d**, Cumulative distribution of mean PTEN intensity in layer 2/3 MGE interneurons. Kruskal–Wallis, $***P = 1.7 \times 10^{-54}$; $n = 223$ cells (P5), 184 cells (P7), 394 cells (P8) and 395 cells (P10) from three animals at each age. **e, f**, Coronal sections through

S1BF from *Pten^{fl/fl}* and *Lhx6-Cre;Pten^{fl/fl}* mice at P16. **g**, Quantification of the density of PV⁺ and SST⁺ cells in P16 *Pten^{fl/fl}* and *Lhx6-Cre;Pten^{fl/fl}* mice. Two-tailed Student's *t*-test, $*P = 0.04$, $**P = 0.005$; $n = 4$ *Pten^{fl/fl}* and 3 *Lhx6-Cre;Pten^{fl/fl}* mice. **h**, Schematic of experimental paradigm. **i**, Coronal sections through layer 2/3 S1BF from P8 *Nex^{Cre/+}* mice injected with hM3Dq-mCherry at P0 followed by vehicle or CNO treatment. PTEN expression is shown as a custom LUT in GABA-masked cells. **j**, Normalized cumulative distribution of PTEN intensity in layer 2/3 S1BF GABAergic cells in vehicle and CNO-treated mice. Kolmogorov–Smirnov test, $***P = 1.05 \times 10^{-111}$; $n = 1,191$ cells (vehicle) and 3,231 cells (CNO) from four vehicle-treated and eight CNO-treated mice. Data are shown as mean \pm s.e.m. Scale bars, 50 μ m (**b, e, f**), 10 μ m (**j**).

than in controls (Fig. 4b, c). Indeed, the fold changes in the density of PV⁺ and SST⁺ interneurons were identical for *Nex^{Cre/+};Bak^{-/-};Bax^{fl/fl}* and *Nkx2-1-Cre;Bak^{-/-};Bax^{fl/fl}* mutant mice (Extended Data Fig. 7b). These results revealed that prevention of pyramidal cell death is sufficient to abolish programmed cell death in MGE and POA interneurons, which reinforces the idea that excitatory input from pyramidal cells onto interneurons during early postnatal development is critical for establishing the appropriate ratio of excitatory and inhibitory cells in the cerebral cortex.

PTEN regulates interneuron cell death

We next investigated the molecular mechanisms through which pyramidal cell activity prevents programmed cell death in cortical interneurons. In the developing nervous system, the serine–threonine

kinase AKT is a critical mediator of neuronal survival^{31,32} that is antagonized by the activity of the phosphatase and tensin homologue PTEN^{33,34}. Consistent with this, we observed that the relative levels of activated AKT (P-AKT/AKT ratio) increased transiently in the neocortex during the period of interneuron cell death (Fig. 5a). Notably, PTEN levels are very heterogeneous among MGE and POA interneurons during the same period (Fig. 5b). PTEN levels were transiently elevated in sparse interneurons in deep and superficial layers of S1, and this increase was concurrent with the peak of interneuron cell death in these layers (Fig. 5c, d and Extended Data Fig. 8a, b). These observations led us to hypothesize that high PTEN levels during this period may drive interneurons towards cell death, and that pyramidal cells might influence this process by regulating PTEN in interneurons.

To test this hypothesis, we generated mice in which we conditionally deleted *Pten* from postmitotic MGE interneurons^{35,36}. We observed that *Lhx6-Cre;Pten^{f/f}* mutant mice had abnormally large jaws and reduced body weight compared to their littermates by P16, probably owing to the embryonic expression of *Lhx6* in the first branchial arch³⁷, which prevented their analysis at later developmental stages. We nevertheless found that *Pten* conditional mutants had a significantly higher density of PV⁺ and SST⁺ interneurons in S1 than control mice (Fig. 5e, f, g and Extended Data Fig. 8c, d), without any difference in relative distribution across layers (Extended Data Fig. 8e). As *Lhx6-Cre* drives recombination in endothelial cells in addition to MGE interneurons³⁵, we examined whether a change in the organization of neocortical blood vessels might contribute to increased survival of interneurons in conditional *Pten* mutants. We found that the density of blood vessels was higher in conditional *Pten* mutants than in controls (Extended Data Fig. 8c, d). However, this change did not affect the density of pyramidal cells (Extended Data Fig. 8c, d), which rules out an indirect effect of blood vessels on interneuron survival through an increase in pyramidal cell density. To rule out a direct effect of blood vessels on interneuron survival, we carried out a second series of experiments using acute pharmacological inhibition of PTEN. We injected the PTEN inhibitor bpV(pic) systemically at P7 and P8 into wild-type mice and analysed blood vessel density in S1 at P10 (Extended Data Fig. 9b, c). Mice injected with the PTEN inhibitor did not exhibit increased blood vessel coverage (Extended Data Fig. 9b, c). By contrast, transient PTEN inhibition during the period of interneuron cell death increased the density of MGE interneurons compared to control mice (Extended Data Fig. 9a, d, e). Mice injected with the PTEN inhibitor outside the normal window of interneuron programmed cell death showed similar densities of PV⁺ and SST⁺ interneurons to controls (Extended Data Fig. 9f–h). These results revealed that PTEN is likely to be required cell-autonomously for interneuron apoptosis during the normal period of interneuron cell death.

Finally, we examined whether pyramidal cell activity influences the survival of interneurons by non-cell-autonomously regulating the expression of PTEN in these cells during the period of interneuron cell death. To this end, we carried out DREADDs experiments similar to those that led to an increased number of cortical interneurons following transient activation of pyramidal cells between P5 and P8 (Fig. 3), but here we analysed PTEN levels in cortical interneurons at P8 (Fig. 5h). We found that PTEN levels were significantly decreased in GABAergic interneurons following the activation of pyramidal cells (Fig. 5i, j). These results strongly suggest that pyramidal cells influence the normal programmed cell death of interneurons through the activity-dependent inhibition of PTEN, which tips the balance between survival and apoptotic signalling pathways in developing interneurons.

Discussion

Our results suggest that programmed cell death in interneurons has evolved as a mechanism responsible for adjusting the final ratio of excitatory and inhibitory neurons in the cerebral cortex, a critical milestone in the assembly of cortical circuits³⁸. Although synaptic mechanisms are known to stabilize excitatory–inhibitory ratios in cortical circuits^{39–41}, this effectively requires that the relative proportions of pyramidal cells and interneurons are within certain parameters^{42–44}. Considering the disproportionate expansion of neocortical areas during human evolution^{45,46}, it is tempting to speculate that the dependency of interneuron survival on pyramidal cells provided an evolutionary advantage for the preservation of appropriate ratios of excitatory and inhibitory cells during the rapid increase in pyramidal cell numbers in the primate lineage.

Our work indicates that interneuron cell death is non-cell-autonomously regulated by pyramidal cells, which seem to be able to rescue connected interneurons from their intrinsically determined cell death¹⁴ by inhibiting the activity of PTEN during a critical window in postnatal development. It is worth noting that a sizable proportion of individuals with autism spectrum disorders (ASD) and macrocephaly

carry deleterious mutations in the *PTEN* gene^{47,48}. Our observations indicate that loss of PTEN function is sufficient to disrupt programmed interneuron cell death, which may in turn alter the cellular balance of excitation and inhibition in the cerebral cortex. This mechanism may contribute to deregulation of cortical information processing and social dysfunction in individuals with ASD who carry PTEN mutations.

The rate of apoptosis in pyramidal cells varies among functionally different neocortical areas and even across layers within the same cortical area⁴⁹. This suggests that the proposed mechanism might sculpt the heterogeneous patterns of interneuron distribution that exist across the cerebral cortex⁵⁰. Consequently, the regulation of programmed cell death in interneurons by pyramidal cells is likely to contribute to the cytoarchitectonical specialization of cortical areas.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0139-6>.

Received: 14 December 2017; Accepted: 6 April 2018;

Published online 30 May 2018.

1. Beaulieu, C. Numerical data on neocortical neurons in adult rat, with special reference to the GABA population. *Brain Res.* **609**, 284–292 (1993).
2. Meyer, H. S. et al. Inhibitory interneurons in a cortical column form hot zones of inhibition in layers 2 and 5A. *Proc. Natl Acad. Sci. USA* **108**, 16807–16812 (2011).
3. Gabbott, P. L. & Somogyi, P. Quantitative distribution of GABA-immunoreactive neurons in the visual cortex (area 17) of the cat. *Exp. Brain Res.* **61**, 323–331 (1986).
4. Hendry, S. H., Schwark, H. D., Jones, E. G. & Yan, J. Numbers and proportions of GABA-immunoreactive neurons in different areas of monkey cerebral cortex. *J. Neurosci.* **7**, 1503–1519 (1987).
5. DeFelipe, J., Alonso-Nanclares, L. & Arellano, J. I. Microstructure of the neocortex: comparative aspects. *J. Neurocytol.* **31**, 299–316 (2002).
6. Fishell, G. & Rudy, B. Mechanisms of inhibition within the telencephalon: “where the wild things are”. *Annu. Rev. Neurosci.* **34**, 535–567 (2011).
7. Marín, O. Interneuron dysfunction in psychiatric disorders. *Nat. Rev. Neurosci.* **13**, 107–120 (2012).
8. Nelson, S. B. & Valakh, V. Excitatory/inhibitory balance and circuit homeostasis in autism spectrum disorders. *Neuron* **87**, 684–698 (2015).
9. Yizhar, O. et al. Neocortical excitation/inhibition balance in information processing and social dysfunction. *Nature* **477**, 171–178 (2011).
10. Hamburger, V. & Levi-Montalcini, R. Proliferation, differentiation and degeneration in the spinal ganglia of the chick embryo under normal and experimental conditions. *J. Exp. Zool.* **111**, 457–501 (1949).
11. Yuan, J. & Yankner, B. A. Apoptosis in the nervous system. *Nature* **407**, 802–809 (2000).
12. Raff, M. C. et al. Programmed cell death and the control of cell survival: lessons from the nervous system. *Science* **262**, 695–700 (1993).
13. Green, D. R. Apoptotic pathways: the roads to ruin. *Cell* **94**, 695–698 (1998).
14. Southwell, D. G. et al. Intrinsically determined cell death of developing cortical interneurons. *Nature* **491**, 109–113 (2012).
15. Verney, C., Takahashi, T., Bhidé, P. G., Nowakowski, R. S. & Caviness, V. S. Jr. Independent controls for neocortical neuron production and histogenetic cell death. *Dev. Neurosci.* **22**, 125–138 (2000).
16. Li, Z. et al. Caspase-3 activation via mitochondria is required for long-term depression and AMPA receptor internalization. *Cell* **141**, 859–871 (2010).
17. Goebbels, S. et al. Genetic targeting of principal neurons in neocortex and hippocampus of NEX-Cre mice. *Genesis* **44**, 611–621 (2006).
18. Xu, Q., Tam, M. & Anderson, S. A. Fate mapping Nkx2.1-lineage cells in the mouse telencephalon. *J. Comp. Neurol.* **506**, 16–29 (2008).
19. Price, D. J., Aslam, S., Tasker, L. & Gillies, K. Fates of the earliest generated cells in the developing murine neocortex. *J. Comp. Neurol.* **377**, 414–422 (1997).
20. Bartolini, G., Ciceri, G. & Marín, O. Integration of GABAergic interneurons into cortical cell assemblies: lessons from embryos and adults. *Neuron* **79**, 849–864 (2013).
21. Ikonomidou, C. et al. Blockade of NMDA receptors and apoptotic neurodegeneration in the developing brain. *Science* **283**, 70–74 (1999).
22. Heck, N. et al. Activity-dependent regulation of neuronal apoptosis in neonatal mouse cerebral cortex. *Cereb. Cortex* **18**, 1335–1349 (2008).
23. Léveillé, F. et al. Suppression of the intrinsic apoptosis pathway by synaptic activity. *J. Neurosci.* **30**, 2623–2635 (2010).
24. Madisen, L. et al. Transgenic mice for intersectional targeting of neural sensors and effectors with high specificity and performance. *Neuron* **85**, 942–958 (2015).
25. Bortone, D. & Polleux, F. KCC2 expression promotes the termination of cortical interneuron migration in a voltage-sensitive calcium-dependent manner. *Neuron* **62**, 53–71 (2009).
26. Priya, R. et al. Activity regulates cell death within cortical interneurons through a calcineurin-dependent mechanism. *Cell Reports* **22**, 1695–1709 (2018).

27. Denaxa, M. et al. Modulation of apoptosis controls inhibitory interneuron number in the cortex. *Cell Reports* **22**, 1710–1721 (2018).
28. Anastasiades, P. G. et al. GABAergic interneurons form transient layer-specific circuits in early postnatal neocortex. *Nat. Commun.* **7**, 10584 (2016).
29. Roth, B. L. DREADDs for neuroscientists. *Neuron* **89**, 683–694 (2016).
30. Lindsten, T. et al. The combined functions of proapoptotic Bcl-2 family members Bak and Bax are essential for normal development of multiple tissues. *Mol. Cell* **6**, 1389–1399 (2000).
31. Dudek, H. et al. Regulation of neuronal survival by the serine-threonine protein kinase Akt. *Science* **275**, 661–665 (1997).
32. Datta, S. R. et al. Akt phosphorylation of BAD couples survival signals to the cell-intrinsic death machinery. *Cell* **91**, 231–241 (1997).
33. Stambolic, V. et al. Negative regulation of PKB/Akt-dependent cell survival by the tumor suppressor PTEN. *Cell* **95**, 29–39 (1998).
34. Backman, S. A. et al. Deletion of Pten in mouse brain causes seizures, ataxia and defects in soma size resembling Lhermitte–Duclos disease. *Nat. Genet.* **29**, 396–403 (2001).
35. Fogarty, M. et al. Spatial genetic patterning of the embryonic neuroepithelium generates GABAergic interneuron diversity in the adult cortex. *J. Neurosci.* **27**, 10935–10946 (2007).
36. Lesche, R. et al. Cre/loxP-mediated inactivation of the murine Pten tumor suppressor gene. *Genesis* **32**, 148–149 (2002).
37. Grigoriou, M., Tucker, A. S., Sharpe, P. T. & Pachnis, V. Expression and regulation of Lhx6 and Lhx7, a novel subfamily of LIM homeodomain encoding genes, suggests a role in mammalian head development. *Development* **125**, 2063–2074 (1998).
38. Isaacson, J. S. & Scanziani, M. How inhibition shapes cortical activity. *Neuron* **72**, 231–243 (2011).
39. Xue, M., Atallah, B. V. & Scanziani, M. Equalizing excitation–inhibition ratios across visual cortical neurons. *Nature* **511**, 596–600 (2014).
40. Burrone, J., O’Byrne, M. & Murthy, V. N. Multiple forms of synaptic plasticity triggered by selective suppression of activity in individual neurons. *Nature* **420**, 414–418 (2002).
41. Maffei, A., Nataraj, K., Nelson, S. B. & Turrigiano, G. G. Potentiation of cortical inhibition by visual deprivation. *Nature* **443**, 81–84 (2006).
42. Butt, S. J. et al. The requirement of Nkx2-1 in the temporal specification of cortical interneuron subtypes. *Neuron* **59**, 722–732 (2008).
43. Cobos, I. et al. Mice lacking Dlx1 show subtype-specific loss of interneurons, reduced inhibition and epilepsy. *Nat. Neurosci.* **8**, 1059–1068 (2005).
44. Glickstein, S. B. et al. Selective cortical interneuron and GABA deficits in cyclin D2-null mice. *Development* **134**, 4083–4093 (2007).
45. Lui, J. H., Hansen, D. V. & Kriegstein, A. R. Development and evolution of the human neocortex. *Cell* **146**, 18–36 (2011).
46. Florio, M. & Huttner, W. B. Neural progenitors, neurogenesis and the evolution of the neocortex. *Development* **141**, 2182–2194 (2014).
47. Butler, M. G. et al. Subset of individuals with autism spectrum disorders and extreme macrocephaly associated with germline PTEN tumour suppressor gene mutations. *J. Med. Genet.* **42**, 318–321 (2005).
48. Buxbaum, J. D. et al. Mutation screening of the PTEN gene in patients with autism spectrum disorders and macrocephaly. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **144B**, 484–491 (2007).
49. Blanquie, O. et al. Electrical activity controls area-specific expression of neuronal apoptosis in the mouse developing cerebral cortex. *eLife* **6**, e27696 (2017).
50. DeFelipe, J. Types of neurons, synaptic connections and chemical characteristics of cells immunoreactive for calbindin-D28K, parvalbumin and calretinin in the neocortex. *J. Chem. Neuroanat.* **14**, 1–19 (1997).

Acknowledgements We thank S. Bae for laboratory support, I. Andrew for management of mouse colonies, V. van den Berghe for help with breeding strategies, S. A. Anderson, N. Kessaris, R. L. Mort and K. A. Nave for mouse lines, N. Flames, C. Houart and M. Maravall for critical reading of the manuscript, and members of the Marín and Rico laboratories for stimulating discussions and ideas. This work was supported by a grant from the Wellcome Trust (103714MA) to O.M. F.K.W. was supported by an EMBO postdoctoral fellowship and is currently a Marie Skłodowska-Curie Fellow from the European Commission under the H2020 Programme. K.B. is a Henry Wellcome Postdoctoral Fellow and O.M. is a Wellcome Trust Investigator.

Author contributions F.K.W., K.B., V.S., and O.M. designed experiments. F.K.W., K.B., A.P. and M.F.-O. carried out stereology quantifications. V.S. performed and analysed in vivo imaging experiments. F.K.W. performed and analysed DREADDs experiments, except for the analysis of PTEN levels, which was carried out by K.B. F.K.W. analysed *Bax/Bak* mutant mice. K.B. performed western blots, examined interneuron PTEN levels and analysed *Pten* mutant mice. F.K.W. performed in vivo pharmacological PTEN inhibition experiments. F.K.W., K.B., V.S., and O.M. wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0139-6>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0139-6>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to O.M.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Animals. All experiments were performed following the guidelines of King's College London Biological Service Unit and in accordance with the European Community Council Directive of November 24, 1986 (86/609/EEC). Animal work was carried out under licence from the UK Home Office in accordance with the Animals (Scientific Procedures) Act 1986. Both male and female mice were used indiscriminately throughout the study. For stereology on pyramidal cells, *Nex^{Cre/+}* mice¹⁹ (provided by K. A. Nave) were crossed with *Fucci2* mice⁵¹ (*RCL^{Fucci2aR/+}*, provided by R. L. Mort). For stereology on MGE and POA interneurons, *Nkx2-1-Cre* mice²⁰ (JAX008661) were crossed with *RCL^{tdTomato/tdTomato}* mice²⁸ (JAX 007909). For in vivo calcium imaging experiments, *Nkx2-1-Cre;RCL^{tdTomato/tdTomato}* mice were crossed with *RCL^{GCaMP6s/+}* mice²⁹ (JAX024106) to generate *Nkx2-1-Cre;RCL^{tdTomato/GCaMP6s}* mice. All DREADDs experiments were conducted in mice obtained by crossing *Nex^{Cre/Cre}* mice with CD1 mice. To prevent pyramidal cells from undergoing programmed cell death, *Bak^{-/-};Bax^{fl/fl}* mice³³ (JAX006329) were crossed with *Nex^{Cre/Cre}* mice and the F1 inter-crossed to obtain *Bak^{+/-};Bax^{fl/fl}* and *Nex^{Cre/+};Bak^{-/-};Bax^{fl/fl}* mutants. For MGE interneurons, a similar breeding scheme used *Nkx2-1-Cre* mice instead. For the quantification of pyramidal cells in *Nex^{Cre/+};Bak^{-/-};Bax^{fl/fl}* mice, these mutants were crossed with *Fucci2* mice to obtain *Nex^{Cre/+};Bak^{-/-};Bax^{fl/fl};RCL^{Fucci2aR/+}* mutants. *Pten^{fl/+}* mice⁴⁰ (JAX006440) were crossed with *Lhx6-Cre* mice³⁹ (provided by N. Kessaris) to generate *Lhx6-Cre;Pten^{fl/+}* mice, and F1 inter-crosses led to the production of *Lhx6-Cre;Pten^{fl/fl}* mutant mice. Mice were obtained from The Jackson Laboratories unless otherwise stated.

Histology. Mice were anaesthetized with an overdose of sodium pentobarbital and transcardially perfused with saline followed by 4% paraformaldehyde (PFA). Brains from pups younger than P6 were post-fixed for 4 h while brains from mice older than P6 were post-fixed for 2 h at 4°C. Brains were sectioned either on a sliding microtome at 30 or 40 µm as previously described³² or on a vibratome at 40 or 60 µm. All primary and secondary antibodies were diluted in PBS containing 0.25% Triton X-100 and 2% BSA. The following antibodies were used: goat anti-CTGF (1:200, Santa Cruz), rabbit anti cleaved-caspase-3 (1:200, Cell Signalling), rabbit anti-dsRed (1:500, Clontech), goat anti-mCherry (1:500, Antibodies-online), rabbit anti-GABA (1:2,000, Millipore), mouse anti-GABA (1:500, Sigma), mouse anti-NeuN (1:500, Millipore) mouse anti-parvalbumin (1:1,000, Swant), rabbit anti-parvalbumin (1:5,000, Swant), rat anti-somatostatin (1:300, Millipore) and rabbit anti-PTEN (1:500, Abcam). We used Alexa Fluor-conjugated secondary antibodies (Invitrogen). For biotin amplification, sections were incubated with biotinylated secondary antibody (1:200, Vector labs), followed by Alexa Fluor-conjugated Streptavidin (1:200, Vector labs). Blood vessels were stained with Isolectin-B4-FITC or Isolectin-B4-Dylight 594 (1:500, Vector labs).

Stereology. The total numbers of pyramidal neurons and MGE interneurons in the cerebral cortex were estimated using the optical disector method³³:

$$N = \frac{\sum Q^- t}{h \times \text{asf} \times \text{ssf}}$$

where $\sum Q^-$ is the total number of cells counted, t the mean section thickness, h the height of the optical disector (17 µm for pyramidal neuron stereology, 18 µm for MGE stereology), adjusting for the guard zones (1 µm) above and below the disector, asf stands for the area sampling fraction and ssf stands for the section sampling fraction (frequency of sampling). An ApoTome (Zeiss) equipped with a motorized stage and colour camera was connected to a computer with Stereo Investigator software (MBF Biosciences). The boundaries of the neocortex were first defined with a 2.5× objective (Zeiss).

For pyramidal neurons, sampling was performed with a 63× objective (Zeiss, numerical aperture (NA) 1.4). The counting frame was set at 15 × 15 µm² and the grid size at 400 × 400 µm². The sampling parameters were as follows: asf = 0.0014, ssf = 0.25 (P2); 0.125 (all other ages). For MGE interneurons, sampling was performed with a 40× objective (Zeiss, NA 1.3). For the entire cortex stereology analysis, the counting frame was set at 125 × 125 µm² and the grid size at 900 × 900 µm² (P2), 1,200 × 1,200 µm² (all other ages). The sampling parameters were as follows: asf = 0.019 (P2); 0.011 (all other ages), ssf = 0.125. For the upper/lower cortical layer stereology analysis, the counting frame was set at 125 × 125 µm² and the grid size at 900 × 900 µm². The sampling parameters were as follows: asf = 0.019, ssf = 0.125. For the stereological analysis of the barrel field, the counting frame was set at 200 × 200 µm² and the grid size at 350 × 350 µm². The sampling parameters were as follows: asf = 0.3265, ssf = 0.125.

In vivo imaging. P6 mice were anesthetized with 2% isoflurane and held in a nose-clip. Isoflurane concentration during surgery was maintained between 1–2% and the body temperature was maintained at 37°C by a heating pad. The scalp was cleaned with betadine and cut open to expose the skull covering the dorsal neocortex. The periosteal tissue, surrounding the skull, was gently scraped with a scalpel. The skull was cleaned with betadine and Ringer's solution. A circular custom-made

metal head-post (Luigs and Neumann) was attached over the left hemisphere with cyanoacrylate glue (Henkel). A thin protective layer of glue was applied over the skull. The glue was allowed to dry for 10 min. Dental cement (Paladur) was used to reinforce the attachment of the head-post to the skull. The mouse was injected with buprenorphine (2 µl/g of a 50 µg/ml solution) and returned to its home cage.

At P7, the animal was anesthetized and head-restrained in a custom-made head holder. A 3-mm craniotomy was opened over the posterior–lateral neocortex. This craniotomy encompassed the primary somatosensory cortex (S1). Care was taken not to damage the dura mater. A circular coverslip (3 mm diameter, Harvard Apparatus) was placed over the craniotomy and its edges were sealed with cyano-acrylate glue and reinforced with dental cement. Following surgery, dexamethasone (5 µl/g of a 38 µg/ml solution) was injected subcutaneously. The animal was allowed to recover for at least 2 h in its home cage, following which we commenced imaging at P7.

Imaging sessions lasted for 40–60 min and we imaged the same field of view for consecutive days in three mice. *TdTomato* and *GCaMP6s*⁵⁴ were excited using a Ti-Sapphire laser (Coherent Chameleon) tuned to $\lambda = 930$ nm. The emitted photons were collected by two GaAsP detectors through a 20× objective (Olympus, 1.0 NA). The field of view (FOV) measured 385 × 385 µm (512 × 512 pixels). The scan speed was set to 30 Hz and image sequences were obtained in sweeps of 1 min (1800 images per channel per min). The average excitation power was between 40 and 50 mW, and this was kept constant over all imaging days.

To correct for motion artefacts, image registration was carried out using custom-written spatial cross-correlation methods on the *tdTomato* channel. In brief, on every 1-min sweep, a part of the *tdTomato* image sequence, where the animal was not moving, was chosen and 20 frames were averaged to give a non-moving reference image. Every frame of the *tdTomato* image sequence was spatially cross-correlated to this reference image and offset along the x - and y -axes to match the cross-correlation peak. The offsets obtained for each *tdTomato* frame were applied to the corresponding *GCaMP6s* frame.

Calcium imaging analysis. Circular ROIs (diameter 20 pixels) were manually drawn around *tdTomato*-expressing cell bodies. The mean *GCaMP6s* fluorescence intensity in time was extracted. Changes in fluorescence signal were calculated as $\Delta F/F_0$, where the baseline fluorescence (F_0) is the mode of a kernel density estimate of F (ksdensity function in Matlab). Calcium events were detected by setting a threshold of 3% change in fluorescence from baseline.

Receiver-operating characteristic curves. To identify whether the calcium event rate (events per min) at P7 could act as a binary classifier in distinguishing cells that would be alive or dead at P8, we plotted the ROC curve⁵⁵ by varying the discrimination threshold (in this case, the P7 event rate) and calculated the AUC. To test for statistical significance, the cell labels were randomly shuffled 5,000 times. On each shuffle, we calculated the ROC curve and the corresponding AUC. We then compared our observed AUC to the distribution of shuffled AUCs. The P value is the fraction of shuffled AUCs \geq observed AUC.

Intracranial injections. *pAAV8-hSyn-DiO-hM3D(Gq)-mCherry* and *pAAV8-hSyn-DiO-hM4D(Gi)-mCherry* were gifts from B. Roth (Addgene plasmids #44361 and #44362)⁵⁶. P0 mice were anaesthetized with isoflurane and mounted in a stereotaxic frame. Pups were injected with 600 nl of virus diluted in PBS and coloured with 0.5% Fast Green (Sigma). Injections were targeted for the somatosensory cortex with an injection rate of 10 nl/s.

Drugs. For DREADDs experiments, CNO (Tocris) was dissolved in 5% dimethyl sulfoxide (Sigma) and then diluted with 0.9% saline to either 0.1 mg/ml or 0.5 mg/ml. Pups were injected with vehicle (0.05% DMSO) or CNO (10 µl per g) subcutaneously for 4 days, twice daily. For the PTEN inhibitor experiments, dipotassium bisperoxovanadium(pic) dehydrate (bpV(pic), Sigma) was dissolved in 0.9% saline to 0.2 mg/ml. Pups were injected with vehicle (0.9% saline) or bpV(pic) (10 µl per g) intraperitoneally for 2 days, twice daily. All treatments for CNO and PTEN inhibitor experiments were randomly assigned.

Western blotting. Mouse somatosensory cortex tissue was homogenized in RIPA lysis buffer containing 50 mM Tris pH 8, 150 mM NaCl, 2 mM EDTA, 0.5% sodium deoxycholate, 0.1% SDS, 1% NP-40 and 1× protease inhibitor cocktail (cOmplete, Sigma). Samples were denatured in Laemmli sample buffer⁵⁷ and run on 10% SDS-PAGE gels. Separated proteins were electrophoretically transferred onto PVDF membranes. Membranes were blocked with 5% BSA in TBST (20 mM Tris-HCl pH 7.5, 150 mM NaCl and 0.1% Tween20) for 1 h and probed with rabbit anti-P-Akt (Ser473, Cell Signalling, 1:1,000) overnight at 4°C, followed by an HRP-conjugated donkey anti-rabbit antibody (Thermo Fisher, 1:10,000). The blots were developed using ECL femto western blotting detection reagents and following read-out, they were stripped (Thermo Fisher). After confirming stripping efficiency, an HRP-conjugated mouse anti-Akt antibody (Cell Signalling, 1:1,000) was added overnight at 4°C. The blots were developed using ECL western blotting detection reagents, the signals were registered and, following stripping, an HRP-conjugated rabbit anti-Actin (Sigma, 1:20,000) was added for 1 h at room temperature. Pico ECL western blotting reagent was used to detect actin levels.

Signals were read on a Li-COR Odyssey Imaging Band and intensities were analysed using ImageStudioLite.

Image acquisition. Images used for analysis were obtained from an ApoTome (Zeiss), epifluorescence microscope (Leica), or SP8 confocal microscope (Leica). ApoTome images were taken using the ApoTome function in Zen2 software. Images obtained with the confocal and epifluorescence microscope were taken using LAS AF software.

Cell counting. Cortical layers were identified on the basis of their distinct histological characteristics. Layer 1 was identified as a sparsely populated cell layer. The border between layers 2/3 and 4 was distinguished by the higher nuclei density of layer 4. Layer 5 was identified as the layer basal to layer 4 and above layer 6, which contains less densely packed nuclei. Cell density, within cortical layers, was quantified either manually or using custom routines written in Matlab (MathWorks). For manual quantification, all analyses were conducted blind and cells were counted in a rectangular area, 551.5 μm wide at the pia surface within the somatosensory cortex, auditory cortex or motor cortex. Cells were counted without using pseudo-colour in Fiji. Automatic quantification was carried out blinded and using morphological operations for image segmentation.

To identify PTEN staining intensity in tdTomato⁺ or GABA⁺ interneurons, self-designed Cell Profiler⁵⁸ pipelines were used. In brief, tdTomato⁺ or GABA⁺ interneurons were identified as primary objects using the global Otsu thresholding method and any objects outside the pre-set diameter range (25–100 pixels) were excluded. PTEN intensity was measured under this cell mask.

Blood vessel analysis. The fraction of the total area covered by blood vessels and the average vessel diameter were quantified blind using 'Vessel Analysis', an ImageJ plugin (http://imagej.net/Vessel_Analysis; N. Govindaraju and M. Elfarnawany).

Statistical analyses. Unless specified, results were plotted and tested for statistical significance using Prism 7. The samples were tested for normality using the Shapiro–Wilk normality test. Unpaired comparisons were analysed using two-tailed unpaired Student's *t*-tests (normally distributed) and Mann–Whitney tests (not normally distributed). Multiple comparisons with single variables were analysed using one-way ANOVA with post hoc Tukey's test (comparing the mean of each column with the mean of every other column) or Dunnett's test (comparing the mean of each column with the mean of a control column) for normally distributed samples. For samples with nonparametric distribution, either Kruskal–Wallis

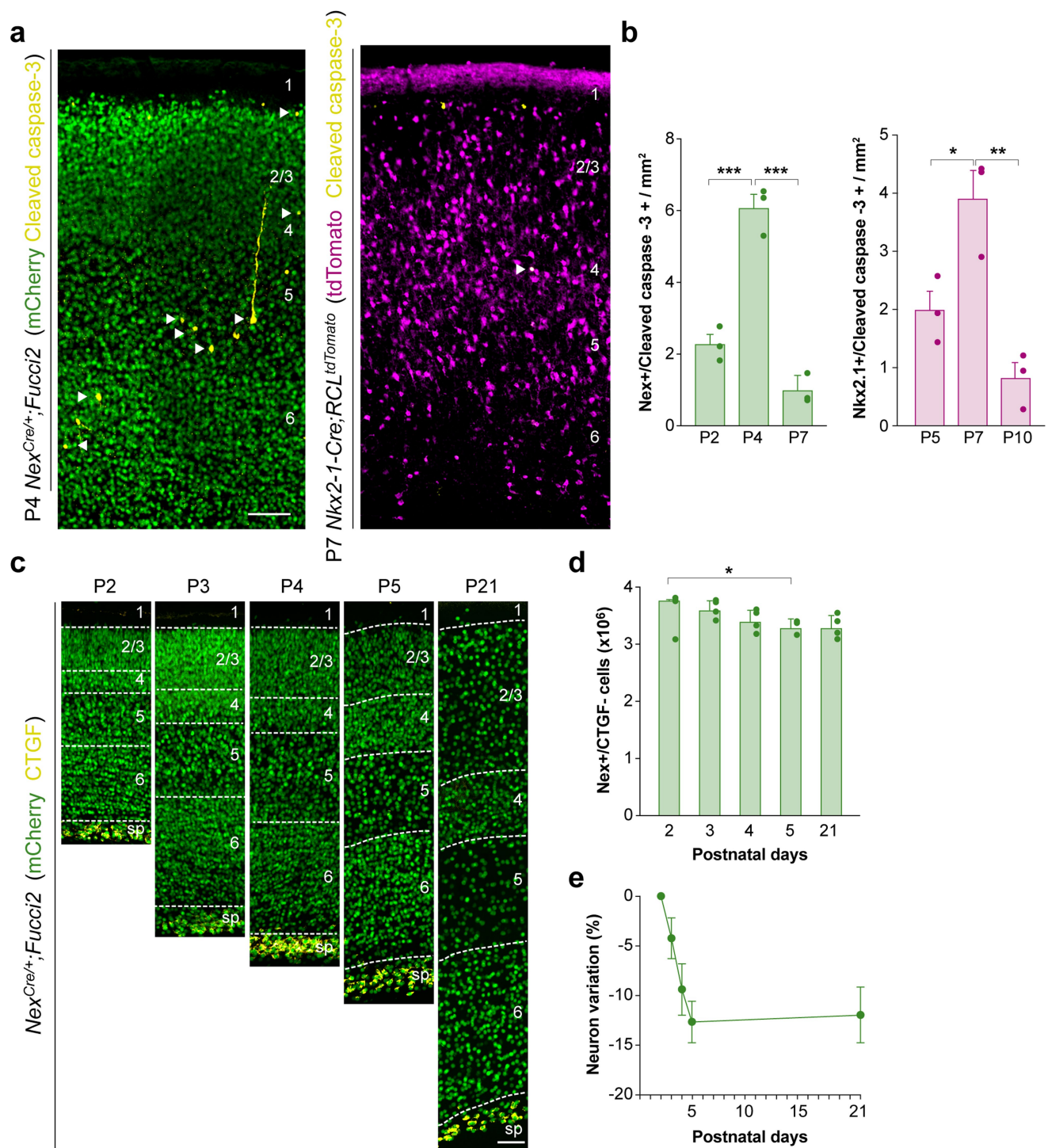
(single measures) or Friedman's test (repeated measures) was performed followed by the post hoc Dunn's test. For multiple comparisons with more than one variable, a two-way ANOVA with post hoc Sidak's test was used. The cumulative distributions of PTEN intensity levels were compared using the Kolmogorov–Smirnov test. Analysis of calcium events rate was carried out in Matlab. In box plots, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers. No statistical methods were used to predetermine sample size. Sample sizes were calculated on the basis of similar published studies. All experiments were replicated at least in two different litters. Unless otherwise stated, the experiments were not randomized (that is, assignments were based on genotypes) and the investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. All data and/or analyses generated during the current study are available from the corresponding author upon reasonable request.

Code availability. For automatic quantification, the code was written in Matlab (Mathworks) and is available from the corresponding author on reasonable request.

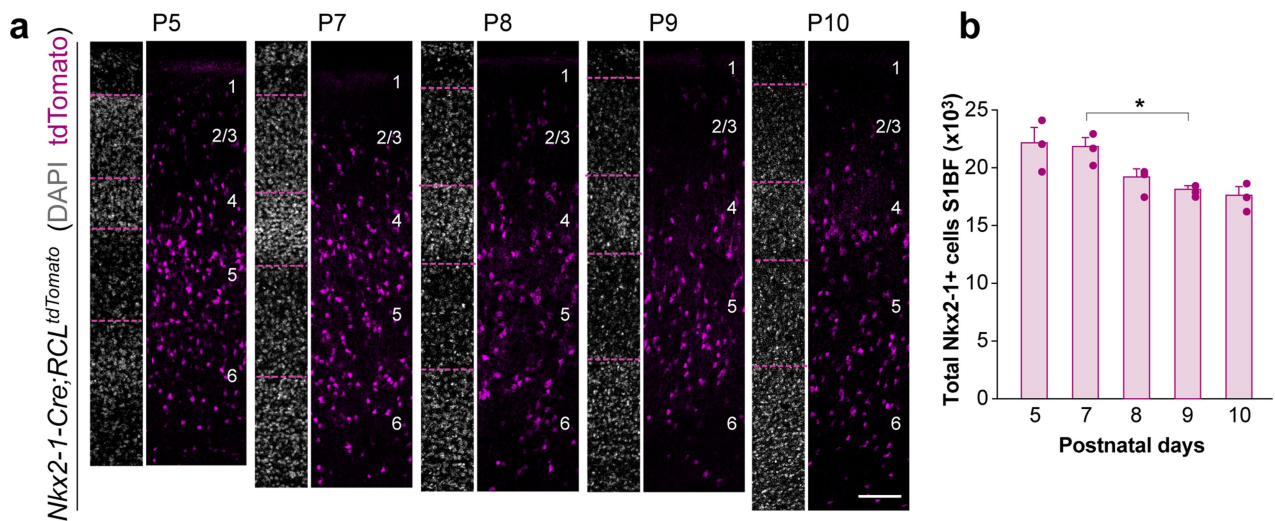
51. Mort, R. L. et al. Fucci2a: a bicistronic cell cycle reporter that allows Cre mediated tissue specific expression in mice. *Cell Cycle* **13**, 2681–2696 (2014).
52. Fazzari, P. et al. Control of cortical GABA circuitry development by Nrg1 and ErbB4 signalling. *Nature* **464**, 1376–1380 (2010).
53. West, M. J. & Gundersen, H. J. Unbiased stereological estimation of the number of neurons in the human hippocampus. *J. Comp. Neurol.* **296**, 1–22 (1990).
54. Chen, T. W. et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
55. Green, D. M. & Swets, J. A. *Signal Detection Theory and Psychophysics*. (Wiley, Oxford, 1966).
56. Krashes, M. J. et al. Rapid, reversible activation of AgRP neurons drives feeding behavior in mice. *J. Clin. Invest.* **121**, 1424–1428 (2011).
57. Laemmli, U. K. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**, 680–685 (1970).
58. Carpenter, A. E. et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**, R100 (2006).



Extended Data Fig. 1 | Extensive cell death in layer 2–6 pyramidal cells.

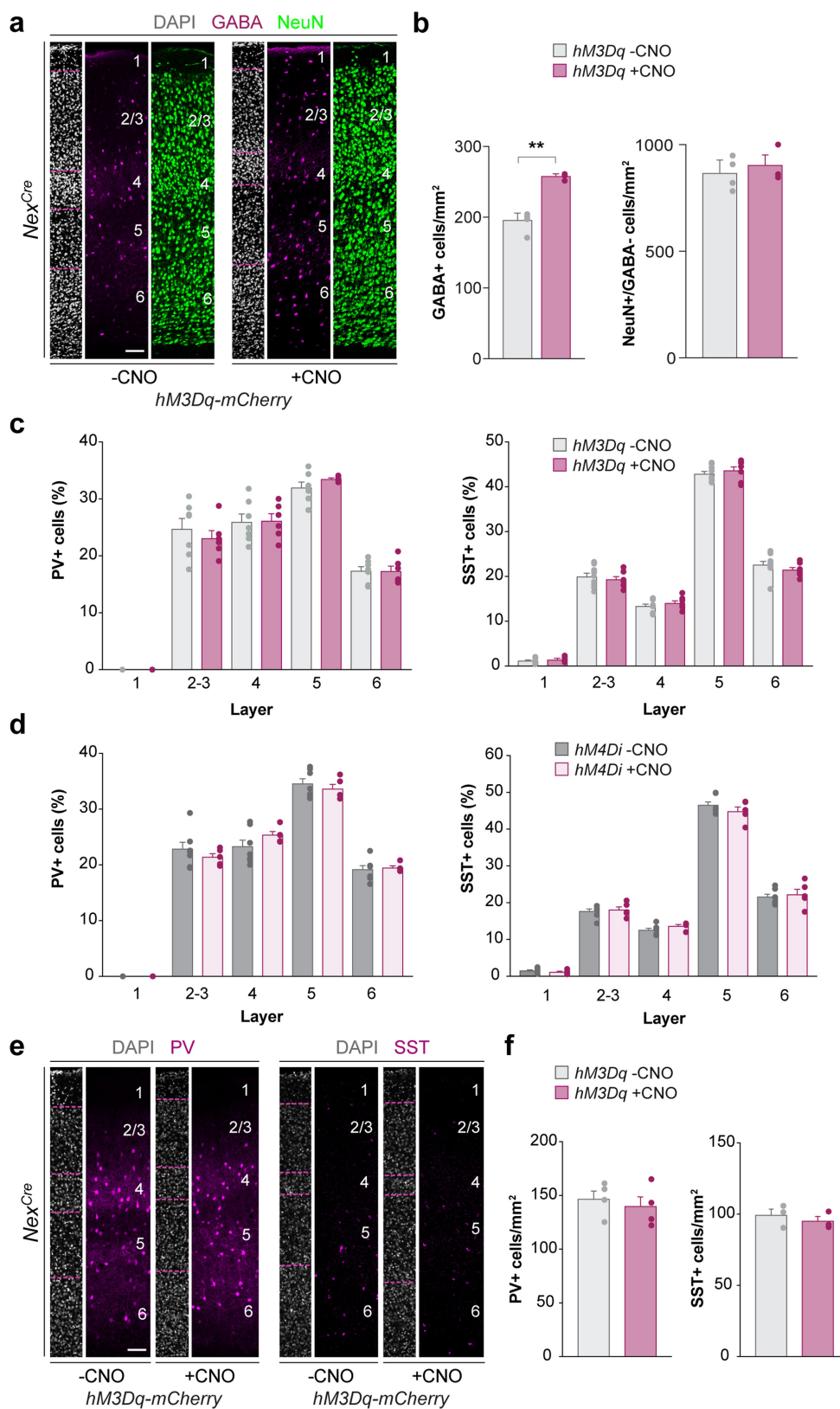
a, Coronal sections through the S1 cortex of P4 *Nex^{Crel/+};Fucci2* (left) and P7 *Nkx2-1-Cre;RCL^{tdTomato}* (right) mice immunostained for cleaved caspase-3 (yellow) and mCherry (green, left) or tdTomato (magenta, right). **b**, Quantification of density of cleaved caspase-3 cells in pyramidal neurons (left, green) and MGE interneurons (right, magenta) during postnatal development (for pyramidal neurons, ANOVA, $F = 73.6$, $***P = 0.003$ (P2 versus P4), $***P = 0.00006$ (P4 versus P7), $n = 3$ mice for all ages; for MGE interneurons, ANOVA, $F = 16.91$, $*P = 0.027$

(P5 versus P7), $**P = 0.0029$ (P7 versus P10), $n = 3$ animals for all ages). **c**, Coronal sections through the barrel cortex of *Nex^{Crel/+};Fucci2* mice during postnatal development immunostained for mCherry (green) and CTGF (yellow). **d**, Total number of pyramidal cells excluding subplate cells in the neocortex of *Nex^{Crel/+};Fucci2* mice (ANOVA, $F = 4.83$ and $*P = 0.03$; $n = 3$ mice for P2 and P5, and 4 mice for P3, P4 and P21). **e**, Temporal variation in the percentage of pyramidal cells excluding the subplate contribution during postnatal development. Data are shown as mean \pm s.e.m. Scale bars, 100 μ m.



Extended Data Fig. 2 | Interneuron cell loss in the barrel field during postnatal development. **a**, Coronal sections through S1BF of *Nkx2-1-Cre;RCL^{tdTomato}* mice (magenta, MGE interneurons) during postnatal development counterstained with DAPI (grey). **b**, Total number of MGE

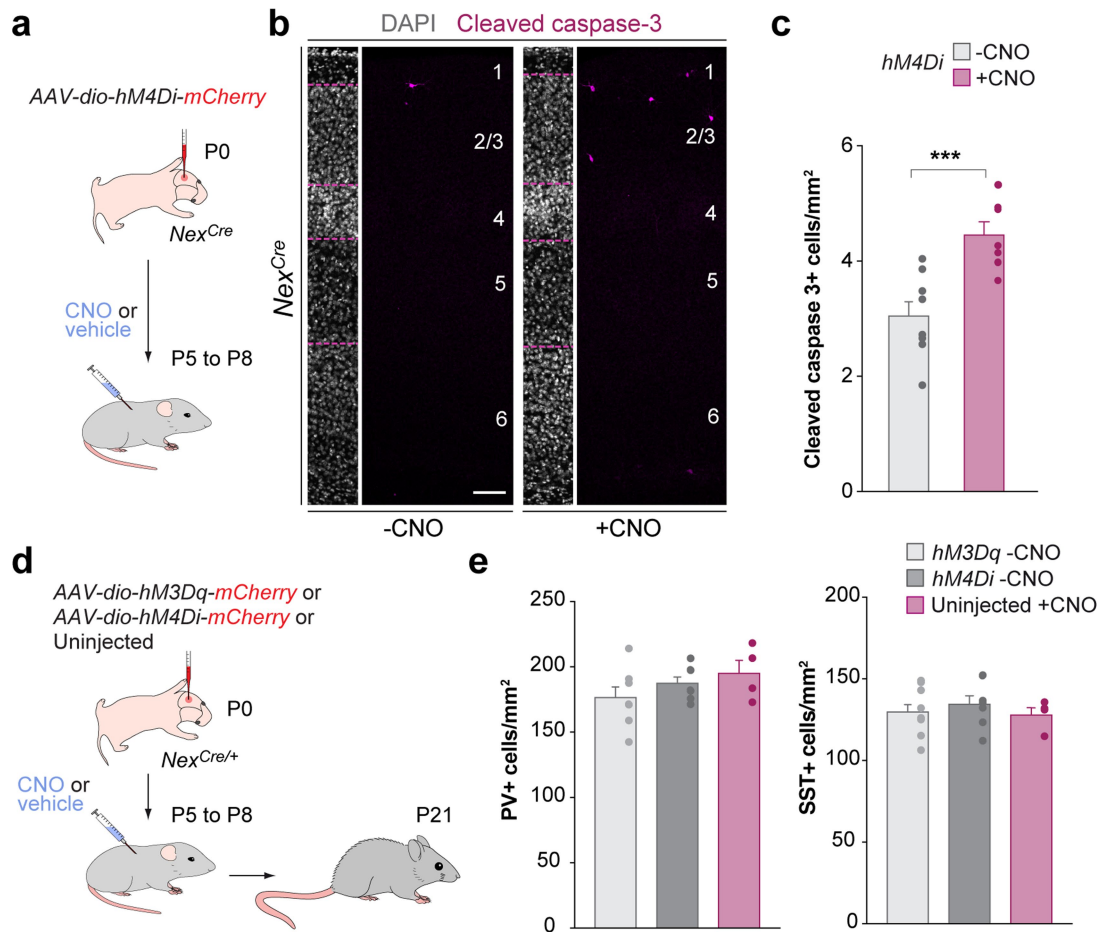
and POA interneurons in S1BF of *Nkx2-1-Cre;RCL^{tdTomato}* mice during postnatal development (ANOVA, $F = 6.40$ and $*P = 0.03$; $n = 4$ animals for each age). Data are shown as mean \pm s.e.m. Scale bar, $100 \mu\text{m}$.



Extended Data Fig. 3 | See next page for caption.

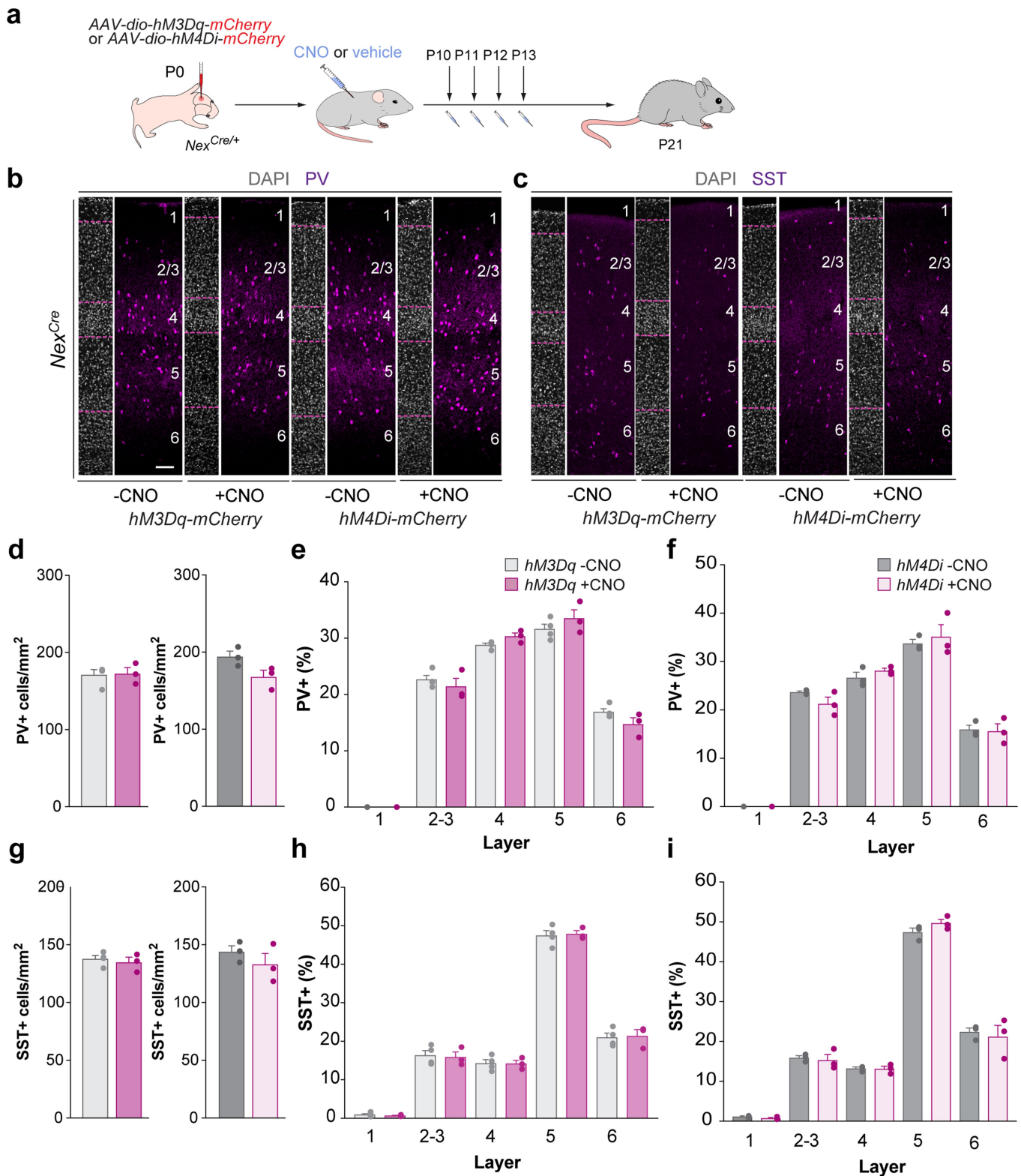
Extended Data Fig. 3 | Alteration of pyramidal cell activity affects interneuron density but not distribution. **a**, Coronal sections through S1BF cortex immunostained for GABA (magenta) and NeuN (green) and counterstained with DAPI (grey) from P21 *Nes^{Cre/+}* mice injected with *hM3Dq-mCherry* virus followed by vehicle or CNO treatment. **b**, Quantification of the density of GABA (left) and NeuN⁺ but GABA⁻ (right) cells in P21 mice injected with *hM3Dq-mCherry* followed by vehicle (grey) or CNO (magenta) treatment (two-tailed Student's unpaired *t*-test, $**P = 0.005$ (GABA), $P = 0.68$ (NeuN⁺ GABA⁻), $n = 4$ animals for vehicle, $n = 3$ animals for CNO conditions). **c**, **d**, Quantification of the distribution of PV⁺ (left) and SST⁺ neurons (right) in P21 *Nes^{Cre/+}* mice injected at P0 with *hM3Dq-mCherry* (**c**) or *hM4Di-mCherry* (**d**) and treated with vehicle (grey) or CNO (magenta) during P5–P8 (two-way ANOVA, $F_{\text{treatment}} = 0.48$, $P = 0.50$ (hM3Dq PV), $F_{\text{treatment}} = -0.04$,

$P = 0.99$ (hM3Dq SST), $F_{\text{treatment}} = 0.88$, $P = 0.37$ (hM4Di PV), $F_{\text{treatment}} = 0.79$, $P = 0.39$ (hM4Di SST); for PV, $n = 7$ animals for hM3Dq and hM4Di – CNO, 6 animals for hM3Dq + CNO, and 5 animals for hM4Di + CNO; for SST, $n = 9$ animals for hM3Dq – CNO, 7 animals for hM3Dq + CNO and hM4Di – CNO, and 5 animals for hM4Di + CNO). **e**, Coronal sections through auditory cortex immunostained for PV (magenta) or SST (magenta) and counterstained with DAPI (grey) from P21 *Nes^{Cre/+}* mice injected with *hM3Dq-mCherry* virus followed by vehicle or CNO treatment. **f**, Quantification of the density of PV⁺ (right) and SST⁺ neurons (left) in auditory cortex in P21 mice injected with *hM3Dq-mCherry* followed by vehicle (grey) or CNO (magenta) treatment (two-tailed Student's unpaired *t*-test, $P = 0.574$ (PV), $P = 0.419$ (SST), $n = 4$ animals for both). Data are shown as mean \pm s.e.m. Scale bars, 100 μm .



Extended Data Fig. 4 | CNO control experiments. **a**, Schematic of experimental design. **b**, Coronal sections through S1 of P8 *NexCre* mice injected with AAV8-dio-hM4Di-mCherry at P0 and treated with (+) or without (–) CNO between P5 and P8, immunostained for cleaved caspase-3 (magenta) and counterstained with DAPI (grey). **c**, Quantification of the density of cleaved caspase-3 cells in P8 mice injected with AAV8-dio-hM4Di-mCherry and treated (magenta) or not treated (grey) with CNO between P5 and P8 (two-tailed Student's unpaired *t*-test, ****P* = 0.009, *n* = 8 animals for –CNO, and *n* = 7

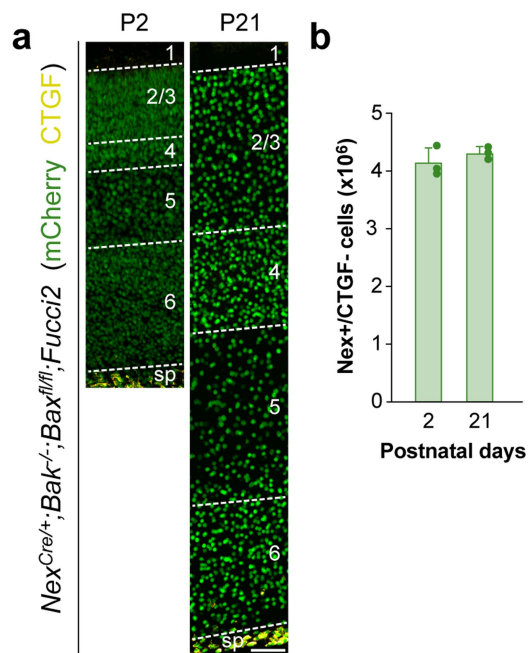
animals for +CNO). **d**, Schematic of experimental design for CNO control experiments. **e**, Quantification of the density of PV⁺ (left) and SST⁺ (right) cells in P21 mice injected with hM3Dq-mCherry or hM4Di-mCherry and not treated with CNO (grey), or not injected with viruses and treated with CNO (magenta) between P5 and P8 (ANOVA, *P* = 0.24 (PV⁺) and *P* = 0.65 (SST⁺); for PV, *n* = 7 animals for hM3Dq and hM4Di –CNO, 4 animals for non-injected +CNO; for SST, *n* = 9 animals for hM3Dq –CNO, 7 animals for hM4Di –CNO, and 4 animals for non-injected +CNO). Data are shown as mean ± s.e.m. Scale bar, 100 μm.



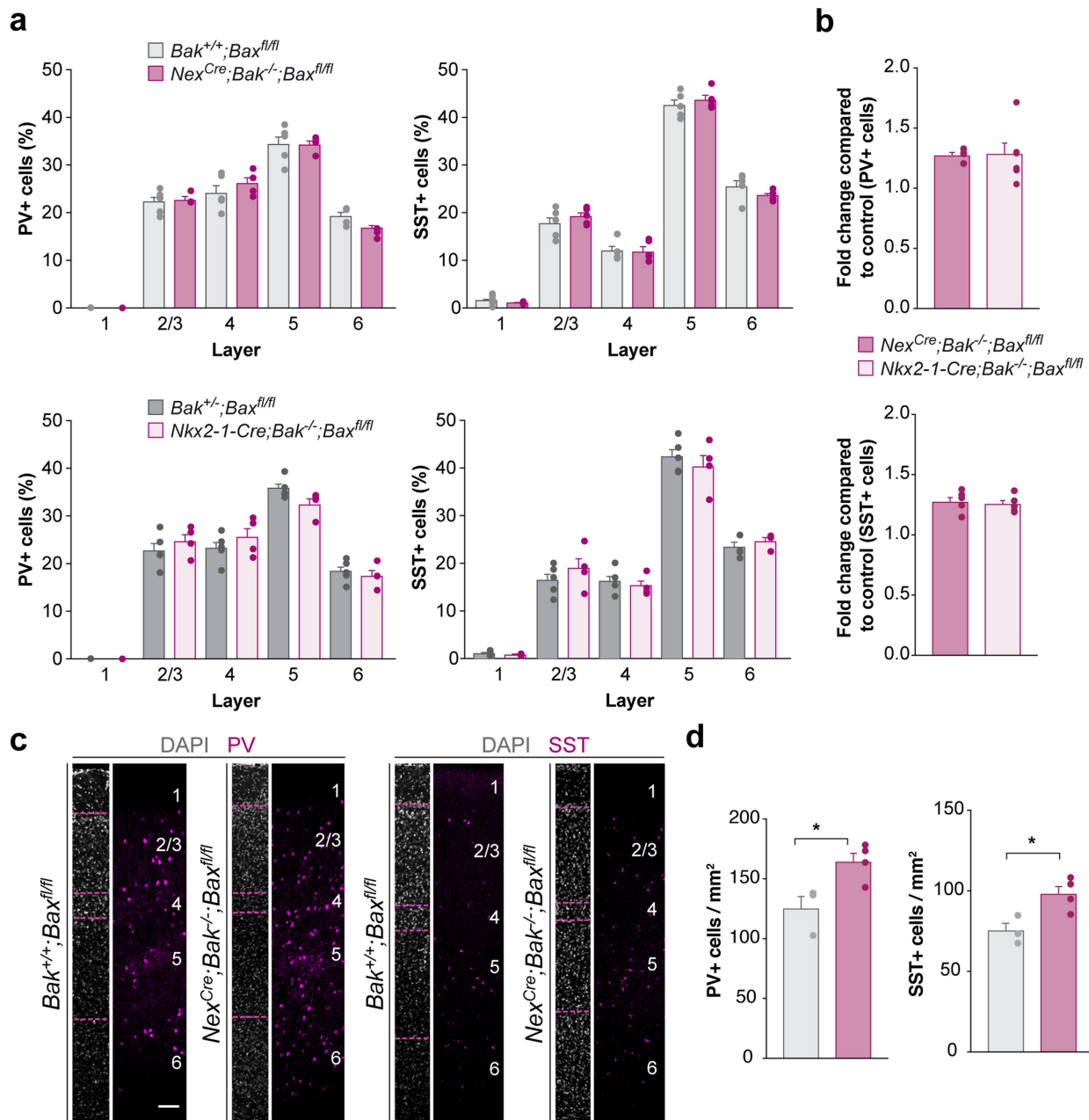
Extended Data Fig. 5 | Alteration of pyramidal cell activity beyond the normal period of interneuron cell death does not affect interneuron survival or distribution. a, Schematic of experimental design.

b, c, Coronal sections through S1BF immunostained for PV (**b**) or SST (**c**) and counterstained with DAPI (grey) from P21 Nex^{Cre/+} mice injected with hM3Dq-mCherry (left) or hM4Di-mCherry (right) viruses followed by vehicle or CNO treatment. **d, g**, Quantification of the density of PV⁺ (**d**) and SST⁺ (**g**) cells in P21 hM3Dq-mCherry injected mice (left bars) and hM4Di-mCherry injected mice (right bars) followed by vehicle (grey bars) and CNO (magenta bars) treatment at P10–P13 (for PV, two-tailed unpaired Student's *t*-test, $P=0.99$ and $P=0.087$, respectively; for SST,

two-tailed unpaired Student's *t*-test, $P=0.56$ and $P=0.37$, respectively; $n=4$ animals for hM3Dq -CNO and 3 animals for all other groups). **e, f, h, i**, Quantification of the distribution of PV⁺ (**e, f**) and SST⁺ cells (**h, i**) in mice injected with hM3Dq-mCherry (**e, h**) or hM4Di-mCherry (**f, i**) followed by vehicle (grey bars) or CNO (magenta bars) treatment at P10–P13 (two-way ANOVA, $F_{\text{treatment}}=0.15$, $P=0.71$ (hM3Dq PV), $F_{\text{treatment}}=0.60$, $P=0.48$ (hM3Dq SST), $F_{\text{treatment}}=1.00$, $P=0.37$ (hM4Di PV), $F_{\text{treatment}}=1.78$, $P=0.25$ (hM4Di SST); $n=4$ animals for hM3Dq -CNO and 3 animals for all other groups). Data are shown as mean \pm s.e.m. Scale bar, 100 μ m.

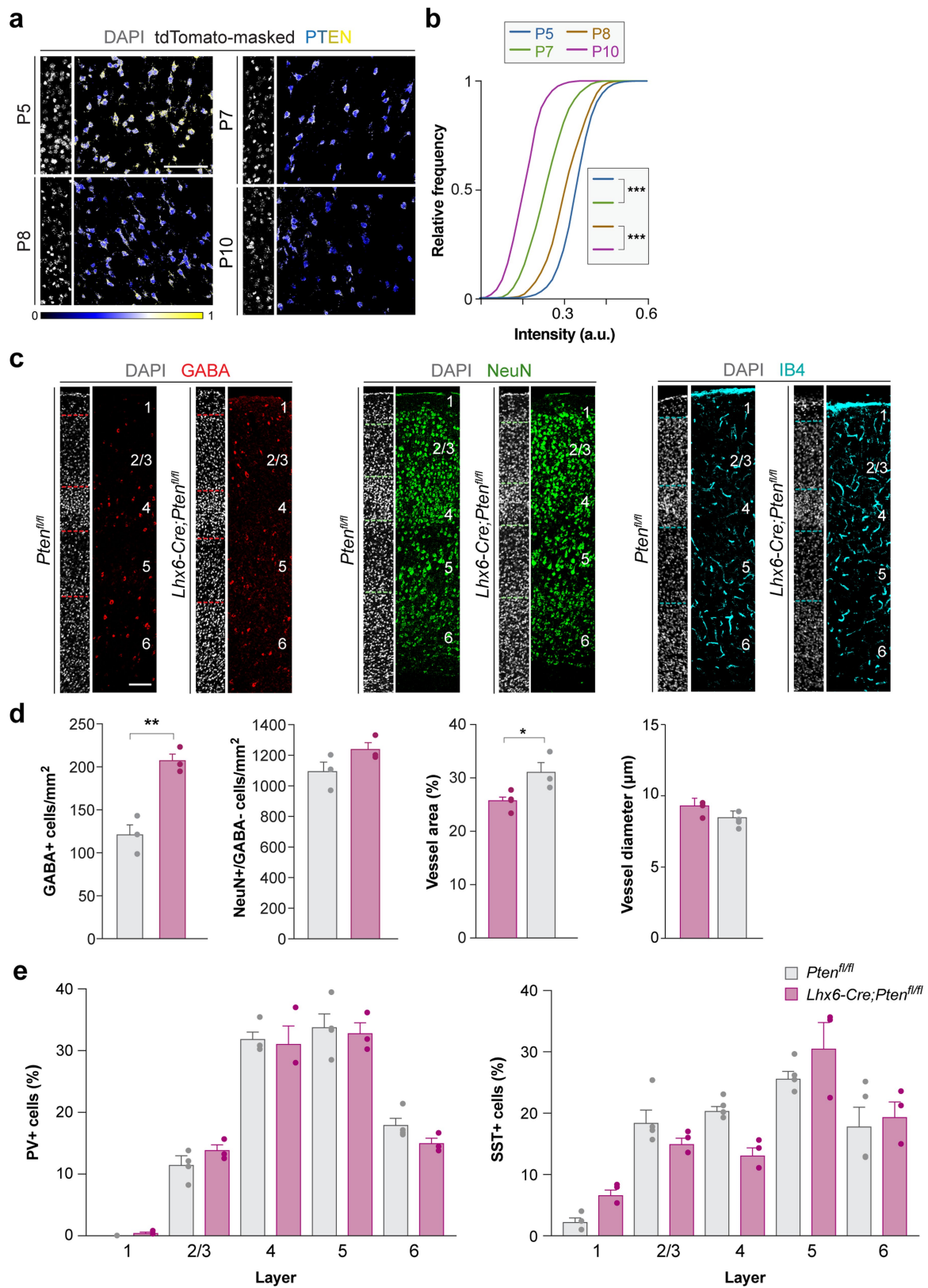


Extended Data Fig. 6 | Loss of BAK and BAX prevents programmed cell death in pyramidal cells. a, Coronal sections through S1BF from P2 and P21 *Nex^{Cre/+}; Bak^{-/-}; Bax^{fl/fl}; Fucci2* mice immunostained for mCherry (green) and CTGF (yellow). **b,** Total number of pyramidal cells (excluding subplate cells) in the neocortex of P2 and P21 *Nex^{Cre/+}; Bak^{-/-}; Bax^{fl/fl}; Fucci2* mice (two-tailed Student's unpaired *t*-test, $P = 0.30$; $n = 3$ animals for both ages). Data are shown as mean \pm s.e.m. Scale bar, 100 μ m.



Extended Data Fig. 7 | Loss of BAK and BAX in pyramidal cells or MGE and POA interneurons affects densities but not lamination of MGE and POA interneurons. **a**, Quantification of the distribution of PV⁺ (left) and SST⁺ (right) interneurons in P30 control (grey), $Nex^{Cre/+};Bak^{-/-};Bax^{fl/fl}$ (dark magenta) and $Nkx2-1-Cre;Bak^{-/-};Bax^{fl/fl}$ (light magenta) mice (two-way ANOVA, $F_{treatment}=3.56$, $P=0.10$ ($Nex^{Cre/+}$ PV), $F_{treatment}=0.44$, $P=0.53$ ($Nkx2-1-Cre$ PV), $F_{treatment}=0$, $P=0.99$ ($Nex^{Cre/+}$ SST), $F_{treatment}=0.44$, $P=0.54$ ($Nkx2-1-Cre$ SST), $n=4$ animals for $Nex^{Cre/+};Bak^{-/-};Bax^{fl/fl}$ (PV) and 5 animals for all other groups). **b**, Quantification of the fold change in the density of PV⁺ (top) and SST⁺ (bottom) interneurons in $Nex^{Cre/+};Bak^{-/-};Bax^{fl/fl}$ (dark magenta) and $Nkx2-1-Cre;Bak^{-/-};Bax^{fl/fl}$ (light magenta) mice compared to their

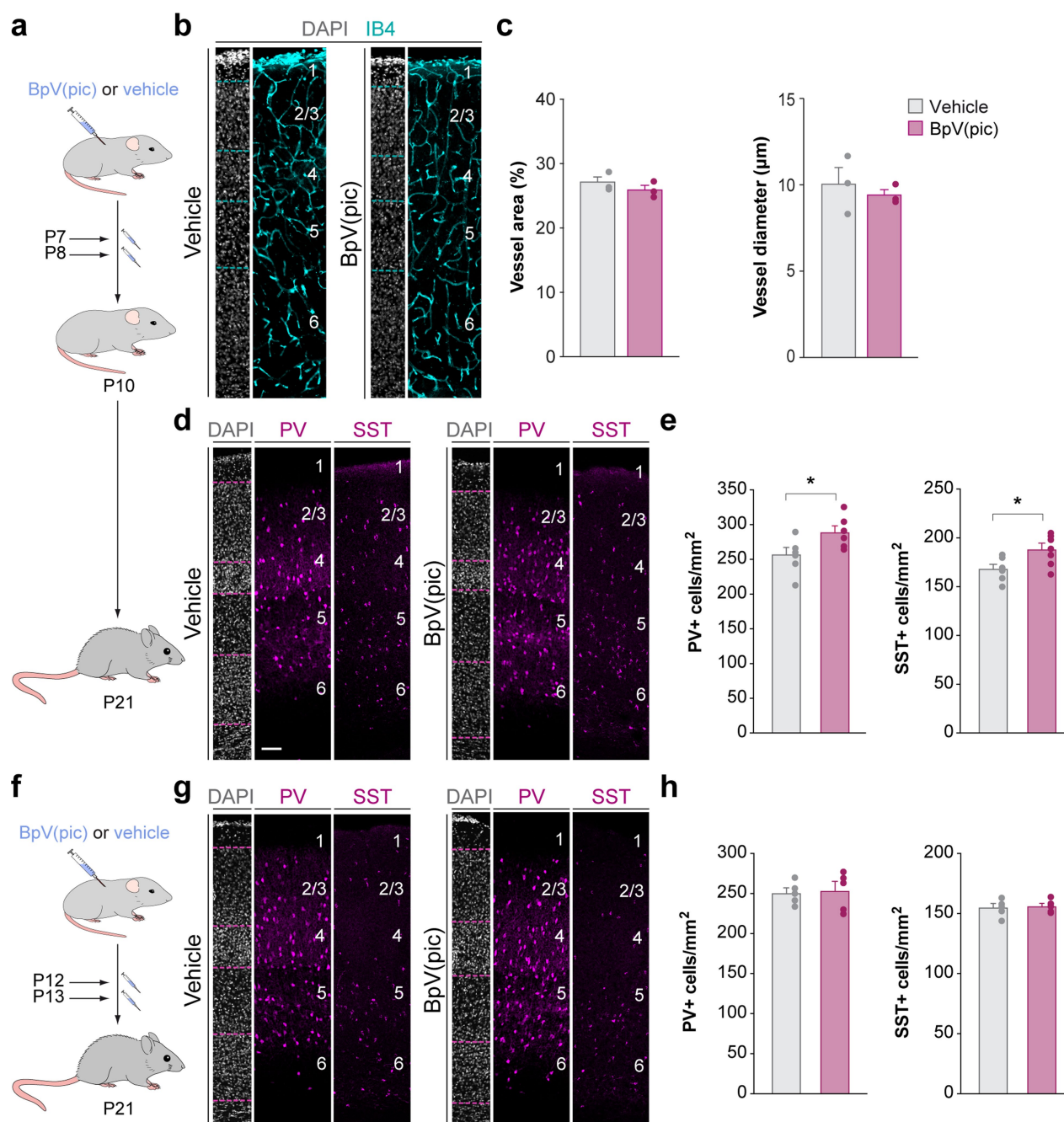
respective controls (two-tailed Student's unpaired t -test, $P=0.90$ (PV), $P=0.67$ (SST); for PV, $n=4$ animals for $Nex^{Cre/+};Bak^{-/-};Bax^{fl/fl}$, 6 animals for $Nkx2-1-Cre;Bak^{-/-};Bax^{fl/fl}$; for SST, $n=5$ animals for both $Nex^{Cre/+};Bak^{-/-};Bax^{fl/fl}$ and $Nkx2-1-Cre;Bak^{-/-};Bax^{fl/fl}$). **c**, Coronal sections through the motor cortex of P30 $Bak^{+/+};Bax^{fl/fl}$ and $Nex^{Cre/+};Bak^{-/-};Bax^{fl/fl}$ mice immunostained for parvalbumin (PV, left) and somatostatin (SST, right) and counterstained with DAPI (grey). **d**, Quantification of the density of PV⁺ (left) and SST⁺ (right) cells in the motor cortex of control and pyramidal cell-specific Bax/Bak double mutant mice at P30 (two-tailed Student's unpaired t -test, * $P=0.02$ (PV), * $P=0.01$ (SST); for PV, $n=4$ animals for both; for SST, $n=3$ animals for both). Data are shown as mean \pm s.e.m. Scale bar, 100 μ m.



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | PTEN expression in deep layer cortical interneurons and effects of loss of PTEN function on neurons and blood vessels. **a**, Coronal sections through layer 5 of S1BF from *Nkx2-1-Cre;RCL^{tdTomato}* mice at P5, P7, P8 and P10, immunostained for PTEN and counterstained with DAPI (grey). PTEN expression is shown as a custom LUT in tdTomato-masked cells. **b**, Cumulative distribution of mean PTEN intensity in layer 5 and 6 MGE and POA interneurons (Kruskal–Wallis test, $***P = 0$; $n = 7,270$ cells (P5), 4,544 cells (P7), 6,780 cells (P8) and 5,043 cells (P10) from 3 mice at each age). **c**, Coronal sections through S1BF from *Pten^{fl/fl}* and *Lhx6-Cre;Pten^{fl/fl}* mice at P16 immunostained for GABA (red, left), NeuN (green, middle) and isolectin B4 (IB4, cyan, right)

and counterstained with DAPI (grey). **d**, Quantification of the density of GABA⁺ (far left) and NeuN⁺ GABA[−] (left) cells and vessel area (right) and diameter (far right) in P16 *Pten^{fl/fl}* (grey) and *Lhx6-Cre;Pten^{fl/fl}* (magenta) mice (two-tailed unpaired Student's *t*-test, $**P = 0.0035$ (GABA), $*P = 0.0326$ (vessel area), $P = 0.0810$ (vessel diameter); Kolmogorov–Smirnov test, $P = 0.1000$ (NeuN⁺ GABA[−] cells), $n = 3$ mice for both genotypes). **e**, Quantification of the distribution of PV⁺ (left) and SST⁺ (right) cells in P16 *Pten^{fl/fl}* (grey) and *Lhx6-Cre;Pten^{fl/fl}* (magenta) mice (two-way ANOVA, $F_{\text{genotype}} = 0.29$, $P = 0.61$ (PV); $F_{\text{genotype}} = 0.0004$, $P = 0.98$ (SST); $n = 4$ *Pten^{fl/fl}* mice and 3 *Lhx6-Cre;Pten^{fl/fl}* mice). Data are shown as mean \pm s.e.m. Scale bars, 100 μm .



Extended Data Fig. 9 | Pharmacological inhibition of PTEN during the interneuron cell death period increases interneuron survival.

a, f, Schematics of experimental design. **b**, Coronal sections through S1BF from P10 mice injected at P7–P8 with vehicle (left) or BpV(pic) (right) stained for isolectin B4 (IB4, cyan) and DAPI (grey). **c**, Quantification of blood vessel area (left) and diameter (right) in P10 mice treated with vehicle (grey) or BpV(pic) (magenta) (Kolmogorov–Smirnov test (vessel area), $P = 0.60$; two-tailed unpaired Student's t -test (vessel diameter), $P = 0.58$, $n = 3$ animals for each group). **d, g**, Coronal sections through

S1BF from P21 mice injected at P7–P8 (**d**) or P12–P13 (**g**) with vehicle (left) or BpV(pic) (right) and immunostained for PV and SST and counterstained with DAPI. **e, h**, Quantification of the density of PV⁺ (left) and SST⁺ (right) cells in S1BF from P21 mice injected at P7–P8 (**e**) or P12–P13 (**h**) with vehicle (grey) or BpV(pic) (magenta) (P7–P8 groups: two-tailed unpaired Student's t -test, $*P = 0.04$ (PV), $*P = 0.03$ (SST); $n = 7$ mice for each group; P12–P13 groups: two-tailed unpaired Student's t -test, $P = 0.84$ (PV), $P = 0.82$ (SST), $n = 5$ animals for each group). Data are shown as mean \pm s.e.m. Scale bars, 100 μ m.

Structural basis of ubiquitin modification by the *Legionella* effector SdeA

Yanan Dong^{1,7}, Yajuan Mu^{1,7}, Yongchao Xie^{1,7}, Yupeng Zhang^{2,7}, Youyou Han¹, Yu Zhou³, Wenhe Wang¹, Zihui Liu¹, Mei Wu⁴, Hao Wang¹, Man Pan⁵, Ning Xu², Cong-Qiao Xu⁶, Maojun Yang², Shilong Fan², Haiteng Deng², Tianwei Tan¹, Xiaoyun Liu⁴, Lei Liu⁵, Jun Li⁶, Jiawei Wang², Xianyang Fang² & Yue Feng^{1*}

Protein ubiquitination is a multifaceted post-translational modification that controls almost every process in eukaryotic cells. Recently, the *Legionella* effector SdeA was reported to mediate a unique phosphoribosyl-linked ubiquitination through successive modifications of the Arg42 of ubiquitin (Ub) by its mono-ADP-ribosyltransferase (mART) and phosphodiesterase (PDE) domains. However, the mechanisms of SdeA-mediated Ub modification and phosphoribosyl-linked ubiquitination remain unknown. Here we report the structures of SdeA in its ligand-free, Ub-bound and Ub-NADH-bound states. The structures reveal that the mART and PDE domains of SdeA form a catalytic domain over its C-terminal region. Upon Ub binding, the canonical ADP-ribosyltransferase toxin turn-turn (ARTT) and phosphate-nicotinamide (PN) loops in the mART domain of SdeA undergo marked conformational changes. The Ub Arg72 might act as a 'probe' that interacts with the mART domain first, and then movements may occur in the side chains of Arg72 and Arg42 during the ADP-ribosylation of Ub. Our study reveals the mechanism of SdeA-mediated Ub modification and provides a framework for further investigations into the phosphoribosyl-linked ubiquitination process.

Ubiquitination is one of the most prevalent protein modifications in eukaryotic cells, regulating a wide array of essential cellular processes^{1,2}. Ubiquitination is carried out by a three-enzyme cascade (E1, E2 and E3), and results in the transfer of ubiquitin (Ub) to a lysine residue of the substrate³. Prokaryotes do not contain the Ub-proteasome system, but a variety of bacterial pathogens adopt intricate mechanisms to influence the host Ub system to support their own survival⁴. *Legionella pneumophila*, the causative agent of a potentially fatal pneumonia known as Legionnaires' disease^{5,6}, can survive and replicate within host cells by creating a vacuole^{7–9}. The biogenesis of the *Legionella*-containing vacuole is based on the approximately 300 *Legionella* substrates (effectors) that are translocated into host cells^{10–13}. Notably, recent studies have identified that the SidE family effectors from *L. pneumophila* could catalyse Ub transfer to several endoplasmic reticulum-associated human Rab GTPases¹⁴ and the endoplasmic reticulum protein reticulon 4 (RTN4)¹⁵, using a unique approach that differs from the canonical ubiquitination pathway¹⁶. By targeting RTN4, the SidE family proteins could control the dynamics of tubular endoplasmic reticulum and promote structural transformations of the tubules¹⁵.

The SidE family members are large proteins (approximately 1,500 residues)⁶ (Extended Data Fig. 1), such as SdeA, which contains a deubiquitinase (DUB) domain (SdeA DUB)¹⁷, a PDE domain (SdeA PDE), an mART domain (SdeA mART) and a C-terminal domain (SdeA CTD) (Fig. 1a). During the ubiquitination process, the R42 residue of Ub is first ADP-ribosylated with NAD⁺ by SdeA mART, and then the phosphodiester bond of the ADP-ribosylated Ub (ADPR-Ub) is cleaved by SdeA PDE to make phosphoribosyl Ub (PR-Ub)^{15,16}. PR-Ub can either remain by itself or be linked via a phosphodiester bond to the hydroxyl group of serine residues in either the substrates or SdeA itself in a reaction catalysed by SdeA PDE. However, structural investigation of the mechanism behind these modifications is required.

Overall structure of SdeA

To understand the mechanism underlying SdeA-mediated Ub modifications, we first solved the crystal structure of a truncated SdeA (amino acids 231–1190, hereafter called SdeA(231–1190)) (Extended Data Table 1) at 3.39 Å resolution. This region of the protein includes SdeA PDE, SdeA mART and a part of SdeA CTD (SdeA pCTD). SdeA(231–1190) exhibited ubiquitination activities that were similar to full-length SdeA (Extended Data Fig. 2a, b). The crystal belonged to the C222₁ space group and two SdeA molecules were found in the asymmetric unit. However, results from the PISA server¹⁸, gel filtration chromatography and analytical ultracentrifuge analysis indicated that SdeA mainly exists as a monomer in solution (Extended Data Fig. 2d, e). Notably, the two SdeA molecules in the asymmetric unit exhibited obvious conformational variations, with a core root mean square deviation (r.m.s.d.) of 1.82 Å among 793 C_α atoms. Superimposition of the two molecules revealed a better alignment of their PDE and mART domains, but prominent differences between their pCTDs, indicating that SdeA pCTD might be flexible in solution (Extended Data Fig. 2f). In the SdeA(231–1190) structure, SdeA mART and SdeA PDE interact primarily through hydrophobic interactions and form a catalytic core (Fig. 1b, c and Extended Data Fig. 3a–d), which sits on top of SdeA pCTD. In contrast to the more conserved PDE and mART domains, SdeA pCTD represented an overall novel fold completely composed of α -helices (Fig. 1b), which could be divided into two subdomains on two sides below SdeA mART (Extended Data Fig. 2g, h).

Structure of the mART domain of SdeA

SdeA mART contains a typical Rossmann fold¹⁹ (Fig. 2a), and exhibits the basic characteristics that are conserved among all known bacterial mART toxins^{20,21}. SdeA mART folds in a two-lobe structure with an N-terminal α -helical lobe (residues 594–758) and a C-terminal

¹Beijing Advanced Innovation Center for Soft Matter Science and Engineering, Beijing Key Laboratory of Bioprocess, College of Life Science and Technology, Beijing University of Chemical Technology, Beijing, China. ²Beijing Advanced Innovation Center for Structural Biology, School of Life Sciences, Tsinghua University, Beijing, China. ³National Institute of Biological Sciences, Beijing, China. ⁴Institute of Analytical Chemistry and Synthetic and Functional Biomolecules Center, College of Chemistry and Molecular Engineering, Peking University, Beijing, China. ⁵Tsinghua-Peking Center for Life Sciences, Department of Chemistry, Tsinghua University, Beijing, China. ⁶Department of Chemistry and Key Laboratory of Organic Optoelectronics & Molecular Engineering of Ministry of Education, Tsinghua University, Beijing, China. ⁷These authors contributed equally: Yanan Dong, Yajuan Mu, Yongchao Xie, Yupeng Zhang. *e-mail: fengyue@mail.buct.edu.cn

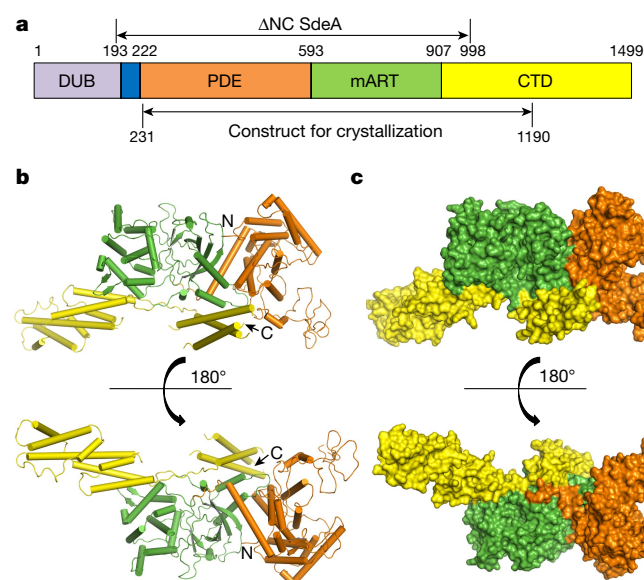


Fig. 1 | Overall structure of SdeA(231–1190). **a**, Domain architecture of SdeA. The residues corresponding to Δ NC SdeA are indicated. **b**, Two different views of the overall structure of SdeA(231–1190) coloured as in **a**. **c**, Two different views of the surface model of SdeA(231–1190).

β -sandwich lobe (Fig. 2a). Notably, structural alignment revealed several unique features in the SdeA mART structure (Extended Data Fig. 4a–d). First, the canonical ARTT and PN loops in SdeA mART are both in different conformations from those of the other mART proteins. Second, the N-terminal α -helical lobe of SdeA mART also shows an overall fold different from those of the others. Third, two consecutive protruding helices are connected by a loop (residues 789–797, here named the plug), which inserts into and interacts with SdeA PDE. This plug loop is, to our knowledge, unique among known mART structures (Fig. 2a and Extended Data Figs. 3a–d, 4a–d). Notably, the much higher ADP-ribosylation activity of SdeA(193–935)^{H277A} when compared to SdeA(597–935) (the single SdeA mART) indicates that unlike SdeA PDE, which is fully active as a single domain (that is, it is able to process ADPR-Ub to PR-Ub and catalyse ubiquitination (Extended Data Fig. 3e)), SdeA mART needs to be stabilized by SdeA PDE to be active. Moreover, deletion of the plug loop of SdeA mART markedly reduced the activity of SdeA(193–998) (Δ NC SdeA)¹⁶ (Fig. 2f and Extended Data Figs. 3f, 4e–h).

SdeA mART has a unique Ub-binding mode

To our knowledge, SdeA mART is the first mART domain ever reported to catalyse ADP-ribosylation of Ub. To understand the molecular mechanism by which SdeA mART recognizes and mediates ADP-ribosylation of Ub, we solved the structure of the SdeA(231–1190)–Ub complex by mixing the two proteins directly in molar ratios of 1:4, 1:6 and 1:8 (SdeA:Ub) before crystallization. Notably, the SdeA(231–1190) structure was solved with three bound Ubs, one at SdeA mART, and the other two at SdeA pCTD (Fig. 2b and Extended Data Fig. 5a–d). The binding of Ub caused a prominent conformational change (r.m.s.d. = 2.24 Å among 830 C α atoms) of SdeA, particularly in the N-terminal region of SdeA pCTD (Extended Data Fig. 5b), again demonstrating the flexibility of SdeA pCTD. However, SdeA pCTD was dispensable for the *in vitro* activity of SdeA, as SdeA(193–935) was fully capable of Ub modification and RAB33B ubiquitination (Extended Data Figs. 4i, 5a). Therefore, we focused on the Ub bound to SdeA mART in the subsequent studies.

To our knowledge, the Ub–SdeA mART binding represents a novel binding mode that differs from all known Ub–protein interactions^{1,22}. Burying a surface area of 607.8 Å², the overall structure of SdeA mART in the complex is similar to that of apo-SdeA mART, with an r.m.s.d.

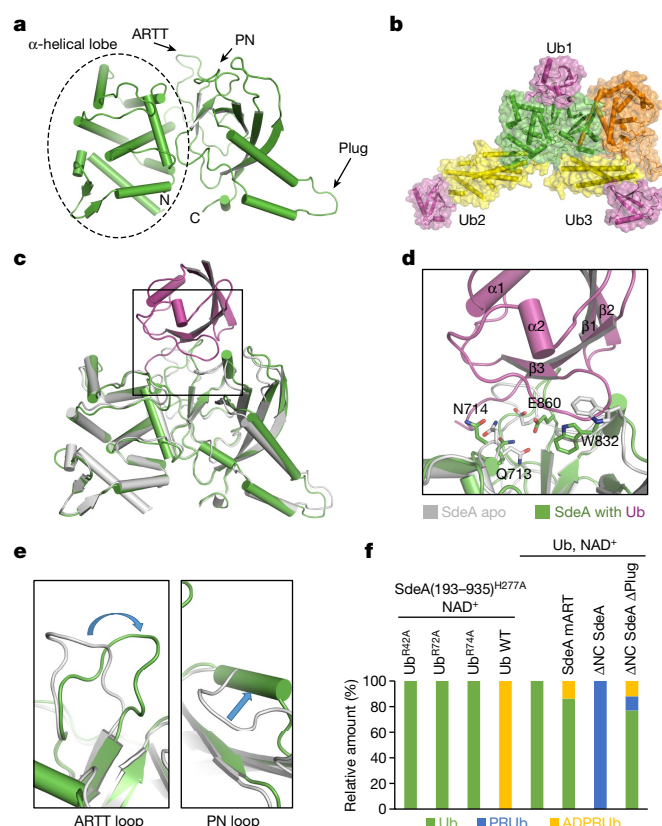


Fig. 2 | SdeA mART undergoes conformational changes upon Ub binding. **a**, Structure of SdeA mART. **b**, Overall structure of the SdeA(231–1190)–Ub complex. SdeA is coloured as in Fig. 1a; three Ub molecules are coloured magenta. **c**, Structural superimposition of apo-SdeA mART (grey) and the SdeA mART–Ub complex (coloured as in **b**) shows conformational changes. **d**, Enlarged view of the outlined region in **c**, major SdeA mART residues which move upon Ub binding are shown in stick representation. **e**, Expanded view showing a superimposition of regions of the ARTT and PN loops before and after Ub binding. The conformational change is highlighted with a blue arrow. **f**, Relative amounts of the unmodified Ub, PR-Ub and ADPR-Ub were studied by top-down liquid chromatography–mass spectrometry analysis after they were isolated from the reaction mixtures. WT, wild type.

of 1.04 Å among 293 C α atoms (Fig. 2c). Nonetheless, the ARTT and PN loops undergo marked conformational changes upon Ub binding (Fig. 2d, e). In response to Ub binding, the C α atom of SdeA^{E860} in the ARTT loop moves towards Ub by 6.0 Å to interact with Ub^{R72} via its side chain, which is simultaneously stabilized by SdeA^{Q713} (Figs. 2d, 3a). In addition to Ub^{R72}, another Ub residue that is essential for its recognition is Ub^{R74}, which inserts into a negatively charged groove on SdeA mART by forming electrostatic interactions with the side chains of SdeA^{D691} and SdeA^{D707} from the α -helical lobe (Fig. 3a and Extended Data Fig. 6a, b). Meanwhile, the C α atom of SdeA^{W832} in the PN loop moves 3.8 Å away from its original position upon Ub binding (Fig. 2d). Notably, SdeA^{W832} acts as a gate for Ub binding, as it occupies the site in the loop that links β 3 and α 2 of Ub in the apo-SdeA structure (Fig. 2d). Together, the ARTT and PN loops and the α -helical lobe of SdeA mART contribute to Ub recognition.

Of the SdeA residues that interact with Ub, SdeA^{E860} has been shown to have an essential role in ADP-ribosylation of Ub^{14,16}. Moreover, SdeA mutant proteins (D691A/D707A, T831D and W832A) all showed decreased activities in Ub phosphoribosylation and substrate ubiquitination, and these mutations also inhibited the yeast toxicity of SdeA (Fig. 3a–c and Extended Data Fig. 7a, b). Consistently, the Ub^{R74A} mutant showed no modification upon treatment with SdeA, as has previously been observed in Ub^{R42A} and Ub^{R72A} mutants¹⁶ (Fig. 3d). We also found that these three Ub mutants were defective in the SdeA

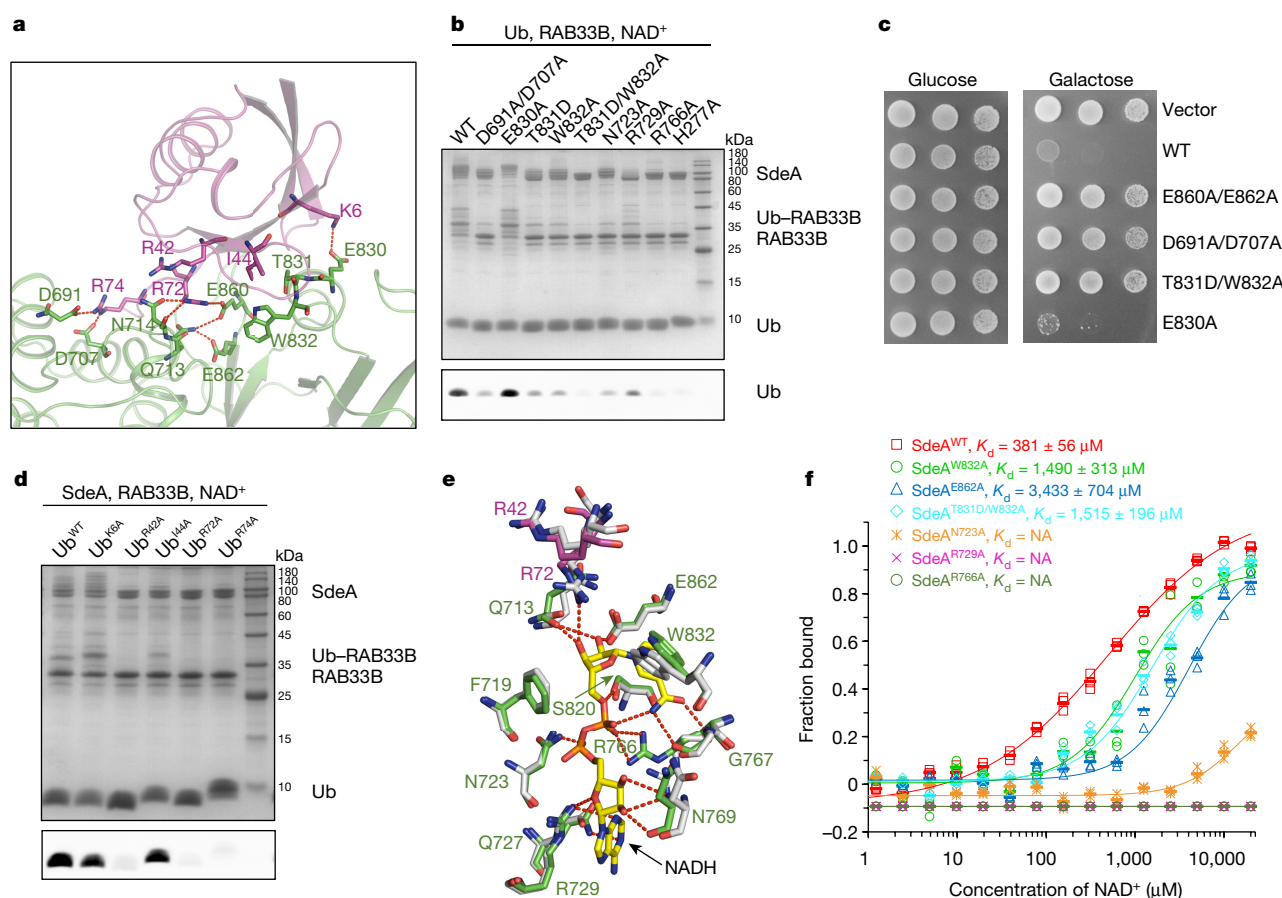


Fig. 3 | Ub^{R72} and Ub^{R74} mediate Ub recognition by SdeA mART.

a, Detailed interactions between Ub (magenta) and SdeA mART (green). Polar interactions are represented by red dashed lines. **b**, Wild-type Δ NC SdeA or the mutant proteins were incubated with Ub, RAB33B and NAD⁺ for 30 min at 37 °C. The samples were then subjected to SDS-PAGE, with Coomassie staining (top) and phospho-specific staining with Pro-Q Diamond (bottom). **c**, Galactose-inducible pYES2 plasmids containing wild-type Δ NC SdeA or the mutants were transformed into yeast W303 strain. Five microlitres of cells in three tenfold serial dilutions were spotted on both glucose- and galactose-containing plates lacking uracil for two days before image acquisition. **d**, Wild-type Ub and the

mART-catalysed reaction, as they could not be ADP-ribosylated by SdeA(193–935)^{H277A} (Fig. 2f and Extended Data Fig. 4f).

The interaction pair of SdeA^{E830} and Ub^{K6}, and the I44 patch of Ub seemed to have a minor role in SdeA mART-Ub binding, as mutations of these residues did not influence Ub modification or substrate ubiquitination (Fig. 3b–d). SdeA mART-mediated ADP-ribosylation of Ub is specific, as Ub-like proteins SUMO1 and NEDD8 could not be ADP-ribosylated by SdeA(193–935)^{H277A}, which is consistent with their structural differences (Extended Data Fig. 6c, d). Structural alignments also suggested that SdeA mART is able to modify both the proximal and distal moieties of K63-, K48-, K11- and M1-linked diubiquitins (Extended Data Fig. 5e–h), a finding that is consistent with a previous study²³.

To further investigate the Ub modification mechanism of SdeA, we solved the structure of SdeA(231–1190)–Ub–NADH complex using a soaking approach (Extended Data Table 1). NADH is reported to be an inhibitor of the mART activity of the enzyme component of the iota-toxin from *Clostridium perfringens*²⁴ and is predicted to occupy the binding site of NAD⁺. Similarly, NADH also had an inhibitory effect on the SdeA-catalysed ubiquitination of RAB33B (Extended Data Fig. 2c). In the structure of the complex, NADH in the cavity forms a ring-like conformation (Fig. 3e and Extended Data Fig. 7e). The overall structure of SdeA mART–Ub remains unchanged upon NADH binding (r.m.s.d.

indicated mutants were incubated with Δ NC SdeA, RAB33B and NAD⁺ for 30 min at 37 °C. The samples were then treated as in **b**. Uncropped blots and gel images for **b** and **d** are shown in Supplementary Fig. 1. **e**, Superimposition of the structures of SdeA–Ub and SdeA–Ub–NADH. The whole SdeA–Ub complex is shown in grey. SdeA and Ub in the SdeA–Ub–NADH complex are shown as in **a**. **f**, MST assays of the binding of NAD⁺ to wild-type SdeA(231–1190) and the indicated mutants. Means are indicated by horizontal lines and individual values from three independent experiments are shown with their markers. Binding curves and K_d values are also shown. NA, for the N723A, R729A and R766A mutants, K_d could not be determined.

of 0.36 Å among 307 C α atoms). Marked conformational changes only occur in the side chains of SdeA^{R766} and SdeA^{W832}, which interact with the phosphate group of NADH through hydrogen bonding and with the nicotinamide ring through π – π interactions, respectively (Fig. 3e). Mutations of the NADH-binding residues of SdeA all impaired the Ub modification, substrate ubiquitination and yeast toxicity of SdeA (Fig. 3b, e and Extended Data Fig. 7c, d, f). The decreased binding to NAD⁺ by the mutant proteins was confirmed using the microscale thermophoresis (MST) assay (Fig. 3f). Notably, although the SdeA variants with mutations in the Ub- or NADH-binding sites were not able to catalyse ubiquitination with normal Ubs, they could utilize and process ADPR–Ub to complete the ubiquitination reaction (Extended Data Fig. 7g–i), indicating that none of the mutations interfere with the activity of SdeA PDE.

Suggested mechanism of SdeA mART ADP-ribosylation

An S_N1 nucleophilic reaction mechanism of the ADP-ribosylation of arginine residues by mART proteins has been proposed by several studies^{25,26}. However, a notable point of the SdeA–Ub–NADH complex structure is the 11.7 Å distance between the nucleophile (NH1/2 atom of Ub^{R42}) and the electrophile (C1D of the N-ribose, the ribose linked to the nicotinamide group of NAD⁺), which is too far for an S_N1 attack (Fig. 4a). Nevertheless, Ub^{R72}, which forms polar interactions with the

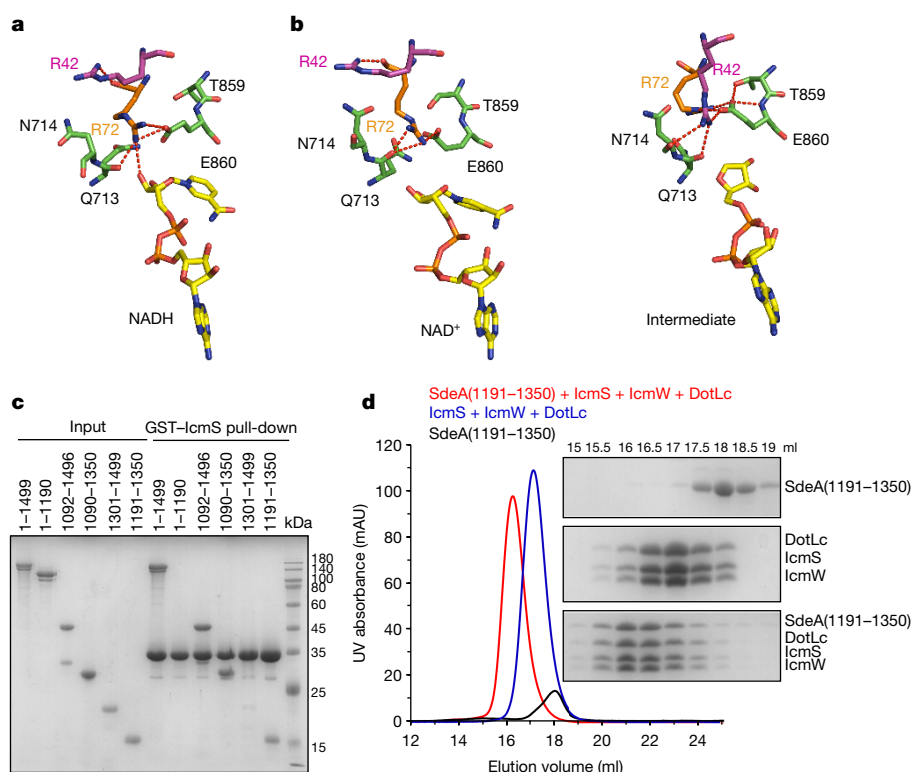


Fig. 4 | Proposed conformational changes during the reaction and function of SdeA CTD. a, The conformations of Ub^{R42} and Ub^{R72} in the SdeA mART-Ub-NADH complex are shown. **b**, Molecular dynamics simulation results of the SdeA mART-Ub-NAD⁺ complex (left) and the SdeA mART-Ub-intermediate system (right). **c**, In vitro glutathione

S-transferase (GST) pull-down assays to detect the IcmS binding region of SdeA. **d**, Superposition of the gel filtration chromatograms of SdeA(1191–1350), the IcmS–IcmW–DotLc complex and a mixture of both. Coomassie blue staining of the peak fractions after SDS–PAGE are shown on the right. **c–d**, Uncropped gel images are shown in Supplementary Fig. 1.

side chain of SdeA^{E860}, the backbone carbonyl of SdeA^{Q713} and the O3D atom of the NADH ribose group, is much closer to the site of the electrophile with a distance of 4.6 Å (Fig. 4a). Because we could not obtain the structure of the SdeA–Ub–NAD⁺ complex, we performed molecular dynamics simulations with NADH replaced by NAD⁺ in the complex, during which the positions of the side chains of Ub^{R42} and Ub^{R72}, and NAD⁺ remained almost unchanged (Fig. 4b and Extended Data Fig. 7j–m).

In the proposed SN₁ mechanism, the highly folded and strained conformation of the nicotinamide mononucleotide region of NAD⁺ induces an equilibrium shift towards formation of an oxocarbenium cation intermediate (NAD⁺ with its nicotinamide group cleaved, hereafter referred to as the intermediate)^{25,26}. Therefore, we also performed molecular dynamics simulations with NADH replaced by the intermediate in the SdeA mART–Ub–NADH structure (Extended Data Fig. 7j–m). Notably, after the simulation, Ub^{R72} moved towards the ARTT loop and away from the intermediate. Instead, Ub^{R42} entered the active site, occupied the original position of Ub^{R72}, and formed electrostatic interactions with SdeA^{E860} (Fig. 4b). After the system reached equilibrium, the average distance between the nucleophile (Ub^{R42}) and the electrophile was 4.46 Å (Fig. 4b and Extended Data Fig. 7m). Consistent with the conformational change observed in molecular dynamics simulation, the conformation of Ub^{R42} is highly variable among the available structures of Ub (Extended Data Fig. 8a).

Together with the structural and biochemical results, we propose that during the catalytic process, Ub^{R72} might function as a ‘probe’, together with Ub^{R74}, by anchoring Ub on SdeA mART. After cleavage of the nicotinamide group from NAD⁺, the strain in the highly folded structure of the intermediate would be alleviated, which might destabilize the binding of Ub^{R72}, causing it to leave. This in turn could facilitate the approach of Ub^{R42} to the active site. However, the exact catalytic cycle still needs further investigation. Moreover, SdeA mART-produced ADPR–Ub will be processed to PR–Ub, and linked to the target protein

by SdeA PDE, the mechanism of which can be gained from the structure of SdeA PDE with its substrates^{27,28}.

SdeA CTD interacts with IcmS–IcmW

Because the N-terminal SdeA DUB and approximately 300 residues of the C-terminal were not included in our crystallized construct, we next studied the solution structure of four constructs of SdeA using the small-angle X-ray scattering (SAXS) method to investigate the spatial position of SdeA PDE–SdeA mART within SdeA (Extended Data Table 2 and Extended Data Fig. 8c–h). Both the scattering profile comparison and the reconstructed molecular envelope indicated that the crystal structure of SdeA(231–1190) is similar to its structure in solution (Extended Data Fig. 8d, e). Superimposing the envelope of SdeA(231–1190) onto the envelope of SdeA(1–1190) further revealed the extra electron density within the envelope of SdeA(1–1190) for SdeA DUB (Extended Data Fig. 8f), which forms a triangle-shaped catalytic core with SdeA mART and SdeA PDE. The envelope of the C-terminal SdeA(1092–1496) indicated the helical-bundle shape of this region (Extended Data Fig. 8g). Further superimposition of the SAXS envelopes of SdeA(1–1499), SdeA(1–1190), SdeA(1092–1496) and the crystal structure of SdeA(231–1190) reconfirmed the positions of SdeA DUB and SdeA CTD (Extended Data Fig. 8h). Notably, in vitro binding assays revealed that SdeA CTD is involved in binding to the adaptor protein complex IcmS–IcmW^{6,29,30} and the minimal binding region is SdeA(1191–1350), which could form a tight complex with IcmS–IcmW–DotLc (residues 656–798 of DotL)³¹ (Fig. 4c, d and Extended Data Fig. 8b). This suggests that SdeA CTD might function in the translocation of SdeA into host cells.

Discussion

An unresolved problem is how the ADPR–Ub is delivered from SdeA mART to SdeA PDE. One possibility is that two or more SdeAs might be close to each other in vivo and ADPR–Ubs produced by one SdeA

mART could be used by the PDE domain of an adjacent one. Our study provides mechanistic insight into the structure and function of SdeA and serves as a foundation for the further studies of phosphoribosyl-linked ubiquitination.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0146-7>.

Received: 24 September 2017; Accepted: 28 March 2018;

Published online 23 May 2018.

- Komander, D. & Rape, M. The ubiquitin code. *Annu. Rev. Biochem.* **81**, 203–229 (2012).
- Yau, R. & Rape, M. The increasing complexity of the ubiquitin code. *Nat. Cell Biol.* **18**, 579–586 (2016).
- Hershko, A., Ciechanover, A. & Varshavsky, A. The ubiquitin system. *Nat. Med.* **6**, 1073–1081 (2000).
- Hubber, A., Kubori, T. & Nagai, H. Modulation of the ubiquitination machinery by *Legionella*. *Curr. Top. Microbiol. Immunol.* **376**, 227–247 (2013).
- Jeong, K. C., Sexton, J. A. & Vogel, J. P. Spatiotemporal regulation of a *Legionella pneumophila* T4SS substrate by the metaeffector SidJ. *PLoS Pathog.* **11**, e1004695 (2015).
- Bardill, J. P., Miller, J. L. & Vogel, J. P. IcmS-dependent translocation of SdeA into macrophages by the *Legionella pneumophila* type IV secretion system. *Mol. Microbiol.* **56**, 90–103 (2005).
- Horwitz, M. A. Formation of a novel phagosome by the Legionnaires' disease bacterium (*Legionella pneumophila*) in human monocytes. *J. Exp. Med.* **158**, 1319–1331 (1983).
- Swanson, M. S. & Isberg, R. R. Association of *Legionella pneumophila* with the macrophage endoplasmic reticulum. *Infect. Immun.* **63**, 3609–3620 (1995).
- Kagan, J. C. & Roy, C. R. *Legionella* phagosomes intercept vesicular traffic from endoplasmic reticulum exit sites. *Nat. Cell Biol.* **4**, 945–954 (2002).
- Luo, Z. Q. & Isberg, R. R. Multiple substrates of the *Legionella pneumophila* Dot/Icm system identified by interbacterial protein transfer. *Proc. Natl Acad. Sci. USA* **101**, 841–846 (2004).
- Zhu, W. et al. Comprehensive identification of protein substrates of the Dot/Icm type IV transporter of *Legionella pneumophila*. *PLoS ONE* **6**, e17638 (2011).
- Lifshitz, Z. et al. Computational modeling and experimental validation of the *Legionella* and *Coxiella* virulence-related type-IVB secretion signal. *Proc. Natl Acad. Sci. USA* **110**, E707–E715 (2013).
- Qiu, J. & Luo, Z. Q. *Legionella* and *Coxiella* effectors: strength in diversity and activity. *Nat. Rev. Microbiol.* **15**, 591–605 (2017).
- Qiu, J. et al. Ubiquitination independent of E1 and E2 enzymes by bacterial effectors. *Nature* **533**, 120–124 (2016).
- Kotewicz, K. M. et al. A single *Legionella* effector catalyzes a multistep ubiquitination pathway to rearrange tubular endoplasmic reticulum for replication. *Cell Host Microbe* **21**, 169–181 (2017).
- Bhogaraju, S. et al. Phosphoribosylation of ubiquitin promotes serine ubiquitination and impairs conventional ubiquitination. *Cell* **167**, 1636–1649. e13 (2016).
- Sheedlo, M. J. et al. Structural basis of substrate recognition by a bacterial deubiquitinase important for dynamics of phagosome ubiquitination. *Proc. Natl Acad. Sci. USA* **112**, 15090–15095 (2015).
- Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).
- Rao, S. T. & Rossmann, M. G. Comparison of super-secondary structures in proteins. *J. Mol. Biol.* **76**, 241–256 (1973).
- Han, S., Arvai, A. S., Clancy, S. B. & Tainer, J. A. Crystal structure and novel recognition motif of Rho ADP-ribosylating C3 exoenzyme from *Clostridium botulinum*: structural insights for recognition specificity and catalysis. *J. Mol. Biol.* **305**, 95–107 (2001).
- Jeong, B. R. et al. Structure function analysis of an ADP-ribosyltransferase type III effector and its RNA-binding target in plant immunity. *J. Biol. Chem.* **286**, 43272–43281 (2011).
- Dikic, I., Wakatsuki, S. & Walters, K. J. Ubiquitin-binding domains—from structures to functions. *Nat. Rev. Mol. Cell Biol.* **10**, 659–671 (2009).
- Puvar, K. et al. Ubiquitin chains modified by the bacterial ligase SdeA are protected from deubiquitinase hydrolysis. *Biochemistry* **56**, 4762–4766 (2017).
- Sakurai, J., Nagahama, M., Oda, M., Tsuge, H. & Kobayashi, K. *Clostridium perfringens* iota-toxin: structure and function. *Toxins (Basel)* **1**, 208–228 (2009).
- Tsurumura, T. et al. Arginine ADP-ribosylation mechanism based on structural snapshots of iota-toxin and actin complex. *Proc. Natl Acad. Sci. USA* **110**, 4267–4272 (2013).
- Tsuge, H. et al. Structural basis of actin recognition and arginine ADP-ribosylation by *Clostridium perfringens* ι -toxin. *Proc. Natl Acad. Sci. USA* **105**, 7399–7404 (2008).
- Akturk, A. et al. Mechanism of phosphoribosyl-ubiquitination mediated by a single *Legionella* effector. *Nature* <https://doi.org/10.1038/s41586-018-0147-6> (2018).
- Kalayil, S. et al. Insights into catalysis and function of phosphoribosyl-linked serine ubiquitination. *Nature* <https://doi.org/10.1038/s41586-018-0145-8> (2018).
- Ninio, S., Zuckman-Cholon, D. M., Cambronne, E. D. & Roy, C. R. The *Legionella* IcmS–IcmW protein complex is important for Dot/Icm-mediated protein translocation. *Mol. Microbiol.* **55**, 912–926 (2005).
- Cambronne, E. D. & Roy, C. R. The *Legionella pneumophila* IcmSW complex interacts with multiple Dot/Icm effectors to facilitate type IV translocation. *PLoS Pathog.* **3**, e188 (2007).
- Kwak, M. J. et al. Architecture of the type IV coupling protein complex of *Legionella pneumophila*. *Nat. Microbiol.* **2**, 17114 (2017).

Acknowledgements We thank B. Zhou for the gift of the yeast W303 strain and the pYES2 vector; Y. Zhang for help in the yeast experiments; J. Ren, X. Zhang and R. Qiao for discussions about the mechanisms of SdeA; S. Zhang for the electron microscopy tests of the SdeA sample; C. Yan for help with the data collection process; the Tsinghua University Branch of China National Center for Protein Sciences Beijing and Y. Xue for providing facility support for NMR analysis of the protein samples; the Protein Chemistry Facility at the Center for Biomedical Analysis of Tsinghua University and W. Zhang for sample analysis; the staff at beamline BL17U1 and BL19U1 of the Shanghai Synchrotron Radiation Facility for their assistance with data collection and X. Zuo at the Advanced Photon Source (APS), Argonne National Laboratory (ANL) and the staff of BL19U2 beamline at the National Center for Protein Science Shanghai and Shanghai Synchrotron Radiation Facility for assistance during data collection. Use of the scattering beamline 12-ID-B resource at APS, ANL is allocated under the GUP-52757 to X.F. This work was supported by the National Key Research and Development Program of China (2017YFA0506500), the National Natural Science Foundation of China (31670766, 21532004, 21475005, and 21622501) and the Fundamental Research Funds for the Central Universities (buctylkxj03).

Reviewer information *Nature* thanks K. Gehring and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Y.F. designed and supervised the project. Y.D., Y.M., Y.H. and W.W. purified the proteins, grew and optimized the crystals, and tested the diffractions of hundreds of crystals. Y.M., Y.X., Y.D., Z.L., Y.H., H.W. and N.X. performed the in vitro activity analysis, MST analysis, yeast toxicity assay and GST pull-down assays. Y.Zho. conducted the molecular dynamics simulations, and C.-Q.X. and J.L. conducted calculations to analyse the results of the molecular dynamics simulations. Y.Zha. and X.F. performed the SAXS analysis of different constructs of SdeA. M.W., H.D. and X.L. performed the mass spectrometric analysis. M.P., T.T. and L.L. contributed to experiment design and helped to supervise the project. J.W., Y.F., M.Y. and S.F. collected and analysed the crystallographic data and solved the crystal structure. Y.F. analysed the data and wrote the paper with the help of all the authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0146-7>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0146-7>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to Y.F.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Protein expression and purification. The full length and various segments of *L. pneumophila* SdeA were amplified by PCR and cloned into pGEX6p-1 or pET22b vectors to produce GST-tagged fusion proteins with a PreScission Protease cleavage site between GST and the target proteins, or His-tagged protein. The SdeA mutants were generated by two-step PCR and were subcloned, overexpressed and purified in the same way as the wild-type protein. The SdeA clone with a deletion of residues 789–797 was made by bridging PCR, and a GSG sequence was added between the two SdeA fragments. The proteins were expressed in *Escherichia coli* strain BL21 and induced using 0.2 mM isopropyl- β -D-thiogalactopyranoside (IPTG) when the cell density reached an OD_{600 nm} of 0.8. For GST-tagged proteins, after growth at 16°C for 12 h, the cells were collected, re-suspended in lysis buffer (1 × PBS, 2 mM dithiothreitol (DTT) and 1 mM phenylmethanesulfonyl fluoride) and lysed by sonication. The cell lysate was centrifuged at 20,000g for 45 min at 4°C to remove cell debris. The supernatant was applied onto a self-packaged GST-affinity column (2 ml glutathione Sepharose 4B; GE Healthcare) and contaminant proteins were removed with wash buffer (lysis buffer plus 200 mM NaCl). The fusion protein was then digested with PreScission protease at 4°C overnight. The protein with an additional five-amino-acid tag (GPLGS) at the N terminus was eluted with lysis buffer. The eluant was concentrated and further purified using a Superdex-200 (GE Healthcare) column equilibrated with a buffer containing 10 mM Tris-HCl pH 8.0, 200 mM NaCl, and 5 mM DTT. The purified protein was analysed by SDS-PAGE. The fractions containing the target protein were pooled and concentrated to 20 mg ml⁻¹. Selenomethionine (Se-Met)-labelled SdeA was expressed in *E. coli* B834 (DE3) cells grown in M9 minimal medium supplemented with 60 mg l⁻¹ Se-Met (Sigma-Aldrich) and specific amino acids: Ile, Leu and Val at 50 mg l⁻¹; Lys, Phe and Thr at 100 mg l⁻¹. The Se-Met protein was purified as described above. The SdeA(231–1190) segment was also cloned into pET22b vector, to make a construct with a C-terminal His tag, which was also used in purification and crystallization.

The fragment of human RAB33B cDNA (residues 1–229) was cloned into the MCS1 of pRSFDuet vector to produce His-tagged fusion protein. The fusion protein was induced in *E. coli* Rosetta (DE3) by 0.2 mM IPTG when the cell density reached an OD_{600 nm} of 0.8. Recombinant His-tagged protein was purified using Ni-affinity column chromatography, ion exchange chromatography and was further subjected to gel filtration chromatography (Superdex-200 column) in buffer containing 10 mM Tris-HCl pH 8.0, 200 mM NaCl, 5 mM DTT.

All Ub mutants used in the ubiquitination assay were cloned into pGEX6p-1 vectors to produce GST-tagged fusion proteins. The proteins were purified according to the protocols for GST-tagged proteins, concentrated to 20 mg ml⁻¹ and stored at –80°C until use. In addition, the GST-tagged wild-type Ub was used in the self-ubiquitination experiments of SdeA(231–1190).

For the IcmS protein and its complexes, IcmS was cloned into pGEX6p-1 vector and purified as stated above. His-tagged IcmS and IcmW were cloned into the MCS1 and MCS2 sites of pRSFDuet vector, respectively. Then the IcmS–IcmW complex was purified as His-tagged proteins according to the protocol described above. For coexpression of IcmS–IcmW–DotLc complex, DotLc, which was cloned into pET22b vector, was co-purified with the bacteria containing the above pRSFDuet vector (His–IcmS and IcmW), according to the same protocol as the IcmS–IcmW complex. For coexpression of the IcmS–IcmW–DotLc–LvgA complex, LvgA was cloned into a modified pET15b vector to produce N-terminal His–MBP tagged LvgA with a PreScission protease digestion site between them. The *E. coli* BL21 (DE3) strain transformed with this vector and the above mentioned pRSFDuet vector expressed the three proteins. DotLc in the above mentioned vector was expressed from the *E. coli* BL21 (DE3) strain. Equal volumes of *E. coli* cultures were co-sonicated and cleared lysate was subjected to Ni-affinity column chromatography. After elution from the column, the protein was treated with PreScission protease, and the complex was purified through ion exchange chromatography and gel filtration chromatography.

Crystallization, data collection and structure determination. The SdeA(231–1190) was concentrated to 20 mg ml⁻¹ in 10 mM Tris-HCl pH 8.0, 200 mM NaCl and 5 mM DTT. Crystals were grown using the hanging-drop vapour diffusion method. Crystals of SdeA were grown at 18°C by mixing an equal volume of the protein (20 mg ml⁻¹) with reservoir solution containing 50% Tacsimate pH 7.0, 0.1 M Tris pH 8.8, 6% sorbitol. The crystals appeared overnight and grew to full size in about 4–5 days. The crystals were cryoprotected in reservoir solution containing 10% glycerol before transfer to liquid nitrogen. Se-Met-labelled protein was crystallized in the same buffer, and the crystals diffracted better than the native crystals. After hundreds of crystal diffraction tests at Micro/Max-007HF from Rigaku and beamlines BL17U1 and BL19U1 of the Shanghai Synchrotron Radiation Facility (SSRF)³², the crystal of the Se-Met-labelled protein suitable for structure determination was finally obtained. Purified C-terminal His-tagged SdeA(231–1190) was mixed with purified Ub in molar ratios of 1:4, 1:6 and 1:8, in which the final concentration of SdeA(231–1190) was 24 mg ml⁻¹. The crystals of the SdeA–Ub complex were also grown at 18°C by mixing an equal volume of the protein

mixture with a reservoir solution containing 0.1 M sodium malonate pH 8.0, 0.1 M Tris pH 8.0, and 30% w/v polyethylene glycol 1,000. The crystals appeared after 2 days and grew to full size in about 4–5 days. These crystals were also used in the NADH-soaking experiment.

All the data were collected at SSRF beamline BL17U1 and BL19U1, integrated and scaled using the HKL2000 package³³. Further processing was carried out using programs from the CCP4 suite³⁴. SHELXD³⁵ was used to locate the positions of selenium sites in SdeA. The identified anomalous scattering sites were input into PHASER³⁶ for single-wavelength anomalous dispersion (SAD) phasing. The real-space constraints were applied to the electron density map in DM³⁷. As the resolution of the data was low, the final model rebuilding of SdeA was performed manually using Coot³⁸ and the apo SdeA structure was refined with PHENIX³⁹ against the SAD data using non-crystallographic symmetry and stereochemistry information as restraints. The structures of the SdeA(231–1190)–Ub and SdeA(231–1190)–Ub–NADH complexes were solved by molecular replacement with the structure of apo-SdeA(231–1190) and Ub (PDB: 1UBQ) as templates. Final Ramachandran statistics: 92.6% favoured, 6.0% allowed and 1.4% outliers for apo-SdeA(231–1190) structure; 93.8% favoured, 4.9% allowed and 1.3% outliers for SdeA(231–1190)–Ub structure; 93.2% favoured, 5.5% allowed and 1.3% outliers for SdeA(231–1190)–Ub–NADH structure. Structural illustrations were generated using PyMOL (v.1.8.0.0, <https://pymol.org/>). Data collection and structure refinement statistics are summarized in Extended Data Table 1.

MST assay. The NAD⁺ affinity of the purified wild-type SdeA(231–1190) and its mutants was measured using the Monolith NT.115 (Nanotemper Technologies). All the proteins used were desalted to MST buffer (10 mM HEPES pH 7.5, 150 mM NaCl) before the experiment. The SdeA proteins were fluorescently labelled according to the manufacturer's procedure and the protein concentration was adjusted to 10 μ M. Then fluorescent dye NT-647-NHS was added, mixed and incubated for 30 min at 25°C in the dark. For each assay, the labelled protein (about 0.1 μ M) was mixed with the same volume of unlabelled NAD⁺ of 16 different serial concentrations at room temperature. The samples were then loaded into premium capillaries (NanoTemper Technologies) and measured at 25°C by using 20% LED power and medium MST power. Each assay was repeated three times. Data analyses were performed using MO.Affinity Analysis v.2.2.4 software. With a confidence of 68%, the K_d value is within the given range.

Structural analysis by SAXS. SAXS measurements were carried out at room temperature at the beamline 12 ID-B of the Advanced Photon Source, Argonne National Laboratory and the beamline BL19U2 of the National Center for Protein Science Shanghai and Shanghai Synchrotron Radiation Facility. The scattered X-ray photons were recorded with a PILATUS 1M detector (Dectris) at 12 ID-B and a PILATUS 100k detector (Dectris) at BL19U2. The setups were adjusted to achieve scattering q values of $0.005 < q < 0.89 \text{ \AA}^{-1}$ (12ID-B) or $0.009 < q < 0.415 \text{ \AA}^{-1}$ (BL19U2), in which $q = (4\pi/\lambda)\sin\theta$, and 2θ is the scattering angle. Thirty 2D images were recorded for each buffer or sample solution using a flow cell, with an exposure time of 0.5–2 s to minimize radiation damage and obtain good signal-to-noise ratio. No radiation damage was observed as confirmed by the absence of systematic signal changes in sequentially collected X-ray scattering images. The 2D images were reduced to 1D scattering profiles using Matlab (12ID-B) or BioXTAS Raw (BL19U2). Scattering profiles of the proteins were calculated by subtracting the background buffer contribution from the sample-buffer profile using the program PRIMUS⁴⁰ according to standard procedures⁴¹. Concentration series measurements (fourfold and twofold dilution and stock solution) for the same sample were carried out to remove the scattering contribution owing to inter-particle interactions and to extrapolate the data to infinite dilution. The forward scattering intensity $I(0)$ and the radius of gyration (R_g) were calculated from the data of infinite dilution at low q values in the range of $qR_g < 1.3$, using the Guinier approximation: $\ln I(q) \approx \ln(I(0)) - R_g^2 q^2/3$. These parameters were also estimated from the scattering profile with a broader q range of $0.006\text{--}0.30 \text{ \AA}^{-1}$ using the indirect Fourier transform method implemented in the program GNOM⁴², along with the pair distance distribution function (PDDF), $P(r)$ and the maximum dimension of the protein, D_{\max} . The parameter D_{\max} (the upper end of the distance r), was chosen so that the resulting PDDF has a short, near-zero-value tail to avoid underestimation of the molecular dimension and consequent distortion in low-resolution structural reconstruction. The Porod volume of solutes (V_{porod}), the volume-of-correlation (V_c), were calculated using the programs PRIMUS and Scatter, respectively. The molecular masses of solutes were calculated on a relative scale using the R_g/V_c power law as previously described⁴³, as well as from AUTOPOROD⁴⁴, independently of protein concentration and with minimal user bias. The theoretical scattering intensity of the atomic structure model was calculated and fitted to the experimental scattering intensity using CRYSol⁴⁵. Low-resolution ab initio shape reconstructions were performed with the program DAMMIN, which generates models represented by an ensemble of densely packed beads⁴⁶, using scattering data within the q range of $0.006\text{--}0.30 \text{ \AA}^{-1}$. Thirty-two independent runs for both programs were performed, and the resulting

models were subjected to averaging by DAMAVER⁴⁷ and were superimposed by SUPCOMB⁴⁸ on the basis of the normalized spatial discrepancy criteria and were filtered using DAMFILT to generate the final model.

Analytical ultracentrifugation. Sedimentation velocity experiments were performed at 20 °C with a Beckman XL-I analytical ultracentrifuge (Beckman Coulter) equipped with a four-cell An-60 Ti rotor. Samples were clarified by centrifugation at 12,000 r.p.m. for 10 min in a tabletop centrifuge before the experiment. Reaction buffer containing 10 mM Tris, pH 8.0, 200 mM NaCl and 5 mM DTT was used as the reference solution, and ~400 µl of SdeA(231–1190) peak fractions ($OD_{280\text{ nm}} = 0.7$) was loaded in two-channel centerpieces fitted with sapphire windows in a four-hole rotor. Absorbance scans were taken at 280 nm versus radial location during centrifugation at 30,000 r.p.m. The differential sedimentation coefficients, $c(s)$, frictional coefficients and molecular mass were calculated using SEDFIT software.

Molecular dynamics simulation. Molecular dynamics simulations were based on the crystal structure of the SdeA–Ub–NADH complex described in this manuscript. The simulations were carried out under two conditions: (1) the NAD⁺-bound complex structure (the SdeA mART–Ub–NADH structure with the NADH displaced by NAD⁺) and (2) the intermediate-bound complex (the SdeA mART–Ub–NADH structure with the nicotinamide group of the ligand removed). Hydrogen atoms were added with the optimal hydrogen-bonding networks and side-chain protonation states determined at pH 7.0 by PROPKA^{49,50}. The protein-chain termini of SdeA were capped with neutral acetyl and methylamide groups. Each system was solvated in a cubic water box with a 10 Å buffer and neutralizing counter ions were added. To mimic experimental assay conditions, a 0.15 M NaCl salt bath was introduced. We used the OPLS-AA 2005 force field⁵¹ parameter set for the protein and ligand, and TIP3P model for water. The parameters for ligands (NAD⁺ and intermediate) were generated from the LigParGen⁵² web server, which applied BOSS software⁵³ to assign the bonded and van der Waals parameters by analogy to the existing atom types. The charges were calculated and assigned by a semi-empirical AM1⁵⁴ calculation using CM1A charge model^{55,56}. Simulations were performed with the Desmond software package⁵⁷. Prepared systems were first minimized using 5,000 steps of a steepest descent algorithm, then equilibrated as follows: the system was heated from 0 to 310 K in the isothermal–isobaric (NpT) ensemble over 100 ps with harmonic restraints of 10.0 kcal·mol^{−1}·Å^{−2} on heavy atoms of protein and ligand, and initial velocities sampled from the Boltzmann distribution. Further equilibration was performed at 310 K with harmonic restraints on the protein and ligand starting at 10.0 kcal·mol^{−1}·Å^{−2} and reduced by 1.0 kcal·mol^{−1}·Å^{−2} in a stepwise fashion every 2 ns, for a total of 20 ns of additional restrained equilibration. Production runs were then made for 200 ns duration in the NpT ensemble. The M-SHAKE algorithm⁵⁸ was applied to constrain all bonds involving hydrogen atoms with a time step of 2 fs. The short-range electrostatic and Lennard–Jones interactions were cut off at 9 Å. Long-range electrostatic interactions were computed by the particle mesh Ewald method⁵⁹.

Yeast toxicity assay. Yeast strain W303 was used for all the experiments. Yeast was grown at 30 °C in YPD for transformation or appropriate selective medium lacking uracil and containing either 2% glucose or galactose as a carbon source. For expression in yeasts, genes were cloned into pYES2 vector containing the galactose-inducible promoter. The integrity of all constructs was verified by sequencing analysis. For each construct, about 1 µg of plasmid DNA was used to transform yeast cells using the standard lithium acetate method. For yeast toxicity experiments, W303 strain cells carrying the defined plasmids were grown overnight in synthetic media lacking uracil and containing 2% glucose. The cells were collected, washed once with sterile water and resuspended in sterile water to an $OD_{600\text{ nm}}$ of 1.0, 0.1 or 0.01. Then 5-µl aliquots of this suspension were spotted onto solid synthetic defined medium lacking uracil and containing either 2% glucose or galactose for protein expression. Plates were grown at 30 °C and images were acquired after 2 days of growth.

Preparation of the ADPR-Ub. For producing ADPR-Ub, 0.1 µM ΔNC SdeA^{H277A} was used in the reaction mixture after tests of enzyme concentrations for the best conversion. In brief, 0.1 µM ΔNC SdeA^{H277A} was incubated with 0.4 mM NAD⁺ and 35 µM Ub at 37 °C for 3 h, after which the reaction mixture was concentrated and loaded onto the Superdex-75 column (GE Healthcare). The peak fractions of Ub were pooled and subjected to mass spectrometric analysis to verify that all the Ubs were in the ADPR-Ub form. ADPR-Ub could be prepared to 100% purity by this method. They were then stored at −80 °C until use.

In vitro ubiquitin-modification and RAB33B-ubiquitination assays. For auto-ubiquitination of SdeA(231–1190) experiments, 6 µM of purified GST-tagged ubiquitin was incubated with 0.9 µM SdeA(231–1190) and RAB33B at 37 °C for 1 h in the presence or absence of 0.1 mM NAD⁺ in a buffer containing 50 mM Tris pH 7.5, 1 mM DTT. After the reaction, the samples were analysed using SDS–PAGE and Coomassie staining. For the RAB33B-ubiquitination experiments, to test whether SdeA(231–1190) was functioning normally, 0.6 µM SdeA, 0.9 µM wild-type SdeA(231–1190), SdeA(231–1190)^{E860A/E862A} or SdeA(231–1190)^{H277A} were incubated with 1 mM NAD⁺ and 35 µM Ub in the presence or absence of

9.5 µM His–RAB33B at 37 °C for 5 min. The samples were then analysed using Tricine gel, Coomassie staining and immunoblotting with anti-His and anti-Ub antibodies. For the ubiquitination experiments of wild-type ΔNC SdeA and the mutants, 1 µM ΔNC SdeA (wild type and mutants) was incubated with 0.1 mM NAD⁺, 35 µM Ub and 7.4 µM His–RAB33B at 37 °C for 30 min. For the NADH-inhibition experiment, 1, 2 or 5 mM NADH was added to the reaction mixture, in which the other components were the same concentrations as above. The samples were then analysed using SDS–PAGE. For phospho-specific staining of PR-Ub, Pro-Q Diamond stain was used according to the manufacturer's instructions. For immunoblotting analysis, primary antibodies were used: anti-His (1:5,000, Transgene, HT501) and anti-Ub (1:500, Santa Cruz Biotechnology, sc-8017). For the time kinetics of the ubiquitination reactions, 1.09 µM ΔNC SdeA (wild type and mutants) or combinations of its fragments were incubated with 1 mM NAD⁺, 35 µM Ub, 9.5 µM His–RAB33B at 37 °C for indicated times. The samples were stained with Coomassie and Pro-Q diamond phosphoprotein stain.

Top-down LC–MS analysis of modified Ub and Ub-like proteins. Wild-type Ub was purchased from Boston Biochem (U-100H), and the untagged Ub mutants were purified from *E. coli* cells. The ubiquitination reactions were performed in a 100-µl system, in which 35 µM Ub, Ub mutants or Ub-like proteins was incubated with specific SdeA fragments or mutants for 2 h in a buffer containing 50 mM NaCl, 50 mM Tris pH 7.5 and 2 mM DTT. The reaction mixtures were then run through 30-kDa molecular mass cut-off filters to obtain the modified Ub or other Ub-like proteins below the filter. The proteins were then subjected to LC–MS analysis. A linear ion-trap mass spectrometer (LTQ Velos Pro, Thermo Scientific) was used for total molecular mass analyses. Liquid chromatography separation was carried out on an EASY-nLC 1200 system (Thermo Scientific). The capillary column (75 µm × 150 mm) with a laser-pulled electrospray tip was home-packed with 4-µm, 100 Å Magic C4AQ silica-based particles (Michrom BioResources). The mobile phase consisted of solvent A (97% H₂O, 3% ACN, and 0.1% FA) and solvent B (20% H₂O, 80% ACN and 0.1% FA). The following gradient was used: solvent B was started at 20% for 3 min and then raised to 50% in 20 min; subsequently, solvent B was rapidly increased to 70% in 2 min and maintained for 20 min before 100% solvent A was used for column equilibration. Eluted peptides from the capillary column were electrosprayed directly onto the mass spectrometer for mass-spectrometry analyses. One full mass-spectrometry scan (m/z 600–1,500) was acquired.

In vitro GST pull-down assay. To detect whether SdeA could bind IcmS alone or its complexes, GST-fused full-length SdeA protein was preloaded on glutathione resins and then incubated with IcmS, IcmS–IcmW, IcmS–IcmW–DotLc or IcmS–IcmW–DotLc–LvgA at 18 °C for 1 h. The samples bound on glutathione resins were washed three times with the washing buffer (50 mM Tris, pH 7.5, 50 mM NaCl) and then analysed by SDS–PAGE and Coomassie blue staining. To detect the region of SdeA responsible for IcmS binding, GST-fused full-length IcmS protein was preloaded on glutathione resins and then incubated with different fragments of SdeA protein at 18 °C for 1 h. The samples were treated as stated above.

Gel-filtration binding assay. The SdeA(1191–1350) and IcmS–IcmW–DotLc complex purified as described above were subjected to gel-filtration analysis (Superdex 200, GE Healthcare). They were mixed at a molar ratio of about 1:1 and incubated at 18 °C for 4 h before gel-filtration analysis in buffer containing 10 mM Tris pH 8.0, 100 mM NaCl. Samples from relevant fractions were applied to SDS–PAGE and visualized by Coomassie blue staining.

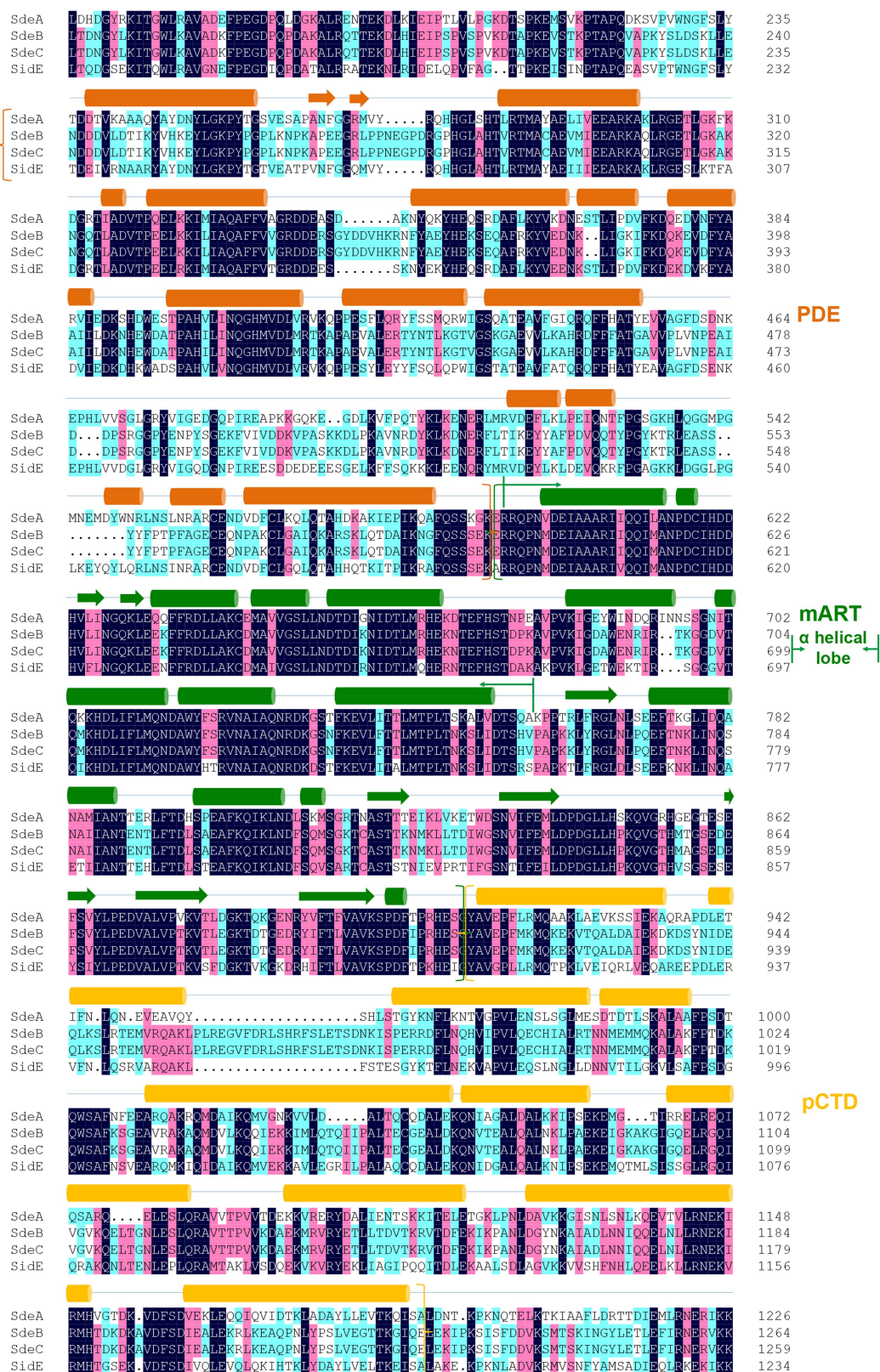
Statistics and reproducibility. No statistical methods were used to predetermine sample size. All of the in vitro assays presented in this work were repeated at least three times with similar results. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. Coordinates and structure factors for the complexes have been deposited in the Protein Data Bank (PDB) under accessions: 5YIM, (SdeA(231–1190)); 5YIK, SdeA(231–1190)–Ub; and 5YIJ, SdeA(231–1190)–Ub–NADH. Uncropped versions of all gels are displayed in Supplementary Fig. 1. All other data are available from the corresponding author upon reasonable request.

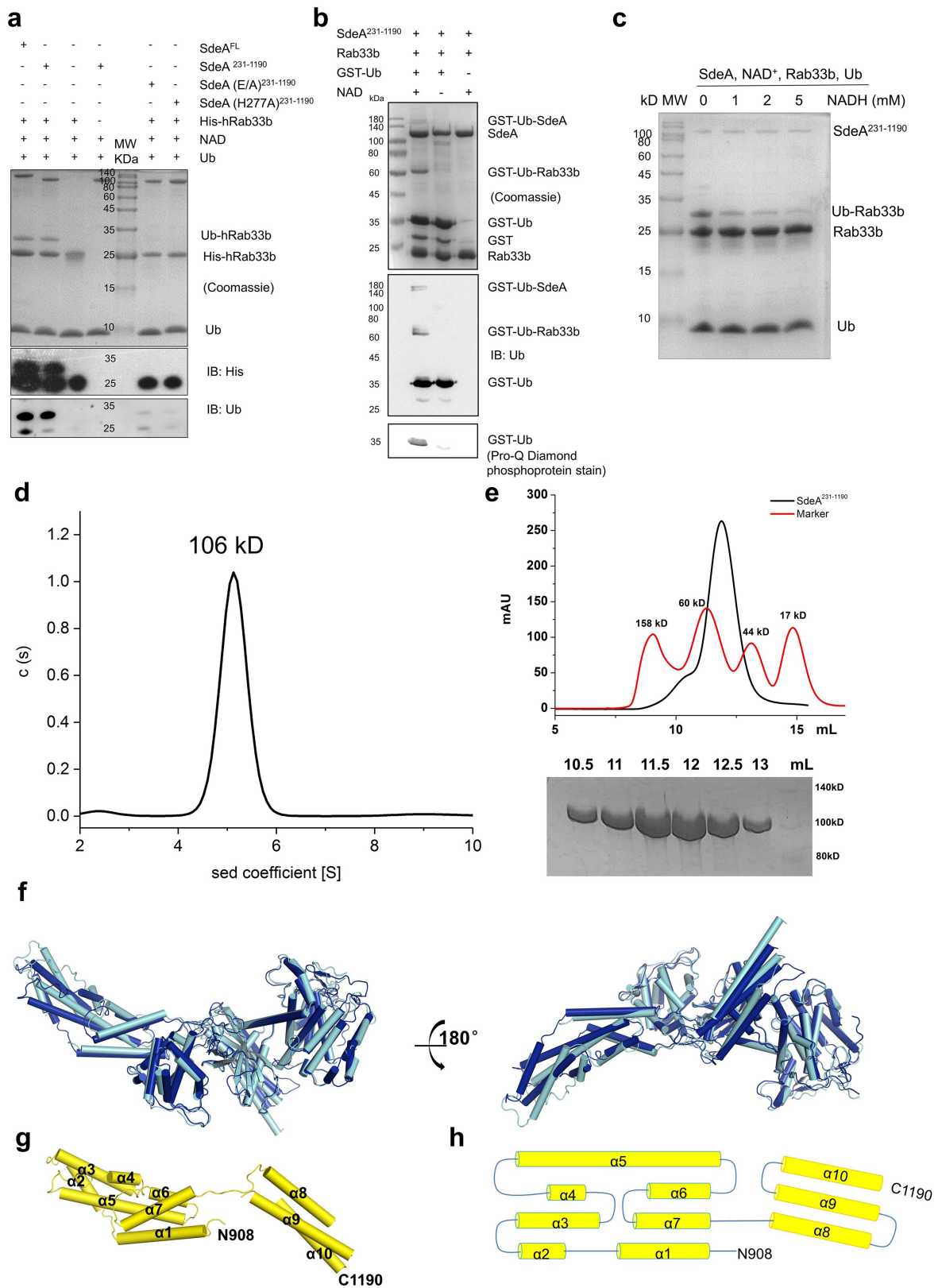
- Wang, Q. S. et al. The macromolecular crystallography beamline of SSRF. *Nucl. Sci. Tech.* **26**, 010102 (2015).
- Otwinowski, Z., Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
- Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
- Schneider, T. R. & Sheldrick, G. M. Substructure solution with SHELXD. *Acta Crystallogr. D* **58**, 1772–1779 (2002).
- McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).

37. Cowtan, K. D. & Zhang, K. Y. Density modification for macromolecular phase improvement. *Prog. Biophys. Mol. Biol.* **72**, 245–270 (1999).
38. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
39. Adams, P. D. et al. PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58**, 1948–1954 (2002).
40. Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. J. & Svergun, D. I. PRIMUS - a Windows-PC based system for small-angle scattering data analysis. *J. Appl. Crystallogr.* **36**, 1277–1282 (2003).
41. Wang, J. et al. A method for helical RNA global structure determination in solution using small-angle X-ray scattering and NMR measurements. *J. Mol. Biol.* **393**, 717–734 (2009).
42. Svergun, D. I. Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *J. Appl. Crystallogr.* **25**, 495–503 (1992).
43. Rambo, R. P. & Tainer, J. A. Accurate assessment of mass, models and resolution by small-angle scattering. *Nature* **496**, 477–481 (2013).
44. Petoukhov, M. V., Konarev, P. V., Kikhney, A. G. & Svergun, D. I. ATSAS 2.1—towards automated and websupported small-angle scattering data analysis. *J. Appl. Crystallogr.* **40**, s223–s228 (2007).
45. Svergun, D. I., Barberato, C. & Koch, M. H. J. CRYSOLE—a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.* **28**, 768–773 (1995).
46. Svergun, D. I. Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophys. J.* **76**, 2879–2886 (1999).
47. Volkov, V. V. & Svergun, D. I. Uniqueness of *ab initio* shape determination in small-angle scattering. *J. Appl. Crystallogr.* **36**, 860–864 (2003).
48. Kozin, M. B. & Svergun, D. I. Automated matching of high- and low-resolution structural models. *J. Appl. Crystallogr.* **34**, 33–41 (2001).
49. Bas, D. C., Rogers, D. M. & Jensen, J. H. Very fast prediction and rationalization of pK_a values for protein–ligand complexes. *Proteins* **73**, 765–783 (2008).
50. Li, H., Robertson, A. D. & Jensen, J. H. Very fast empirical prediction and rationalization of protein pK_a values. *Proteins* **61**, 704–721 (2005).
51. Banks, J. L. et al. Integrated modeling program, applied chemical theory (IMPACT). *J. Comput. Chem.* **26**, 1752–1780 (2005).
52. Dodda, L. S., Cabeza de Vaca, I., Tirado-Rives, J. & Jorgensen, W. L. LigParGen web server: an automatic OPLS-AA parameter generator for organic ligands. *Nucleic Acids Res.* **45**, W331–W336 (2017).
53. Jorgensen, W. L. & Tirado-Rives, J. Molecular modeling of organic and biomolecular systems using BOSS and MCPRO. *J. Comput. Chem.* **26**, 1689–1700 (2005).
54. Storer, J. W., Giesen, D. J., Cramer, C. J. & Truhlar, D. G. Class IV charge models: a new semiempirical approach in quantum chemistry. *J. Comput. Aided Mol. Des.* **9**, 87–110 (1995).
55. Udier-Blagović, M., Morales De Tirado, P., Pearlman, S. A. & Jorgensen, W. L. Accuracy of free energies of hydration using CM1 and CM3 atomic charges. *J. Comput. Chem.* **25**, 1322–1332 (2004).
56. Dodda, L. S., Vilseck, J. Z., Tirado-Rives, J. & Jorgensen, W. L. 1.14*CM1A-LBCC: Localized Bond-Charge Corrected CM1A Charges for Condensed-Phase Simulations. *J. Phys. Chem. B* **121**, 3864–3870 (2017).
57. Bowers, K. J. et al. Scalable algorithms for molecular dynamics simulations on commodity clusters. In *Proc. of the 2006 ACM/IEEE Conference on Supercomputing* 43–43 (2006).
58. Kräutler, V., Van Gunsteren, W. F. & Hünenberger, P. H. A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *J. Comput. Chem.* **22**, 501–508 (2001).
59. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: an $N\log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089 (1993).
60. Petoukhov, M. V. et al. New developments in the ATSAS program package for small-angle scattering data analysis. *J. Appl. Crystallogr.* **45**, 342–350 (2012).



Extended Data Fig. 1 | Sequence alignment of SdeA family members, spanning the regions corresponding to the SdeA fragment used in crystallization. Residues with 100% homology, over 75% homology and over 50% homology are shaded in dark blue, pink and light blue,

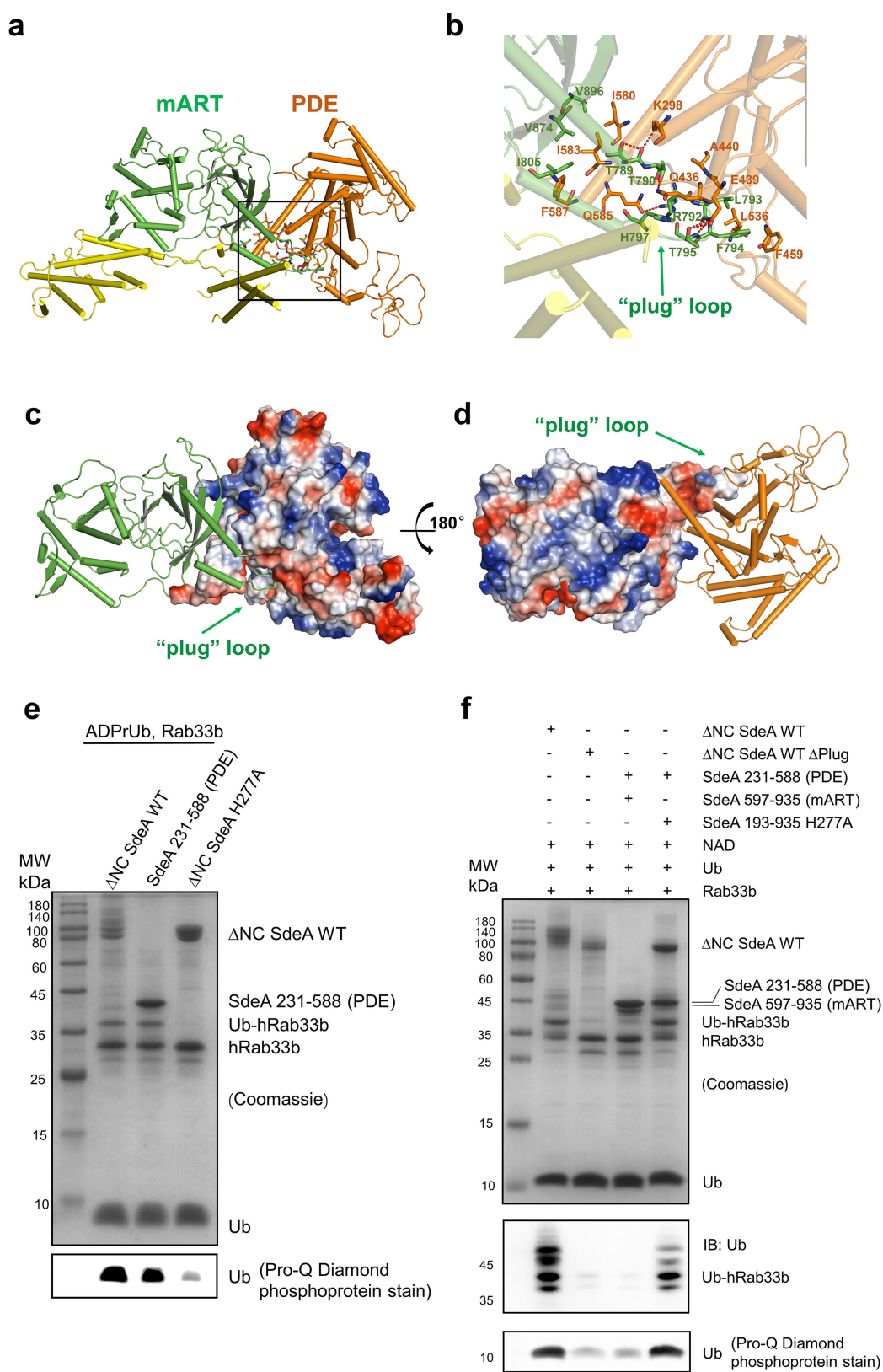
respectively. Secondary structural elements of SdeA are shown above the sequences. The residue ranges of the PDE, mART, pCTD and the α -helical lobe of the mART domain are marked with brackets or lines.



Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | SdeA(231–1190) is an active monomer in solution. **a**, SdeA, wild-type SdeA(231–1190), SdeA(231–1190)^{E860A/E862A} or SdeA(231–1190)^{H277A} were incubated with NAD⁺ and Ub in the presence or absence of His–RAB33B. Ubiquitinated His–RAB33B were analysed using tricine gels, Coomassie staining and immunoblotting with anti-His and anti-Ub antibodies. **b**, SdeA(231–1190) and RAB33B(15–202) were incubated with GST–Ub and NAD⁺, and self-ubiquitinated SdeA was detected by Coomassie staining, immunoblotting with anti-Ub antibodies, and Pro-Q diamond phosphoprotein staining. **c**, SdeA(231–1190), NAD⁺ and RAB33B were incubated with 0, 1, 2 or 5 mM NADH. The ubiquitination reactions were analysed using tricine gel and Coomassie staining. **d**, Analytical ultracentrifugation results showed that SdeA(231–1190) is a monomer. Analytical ultracentrifugation analysis yielded a sedimentation coefficient of 5.13 S, and a molecular mass of

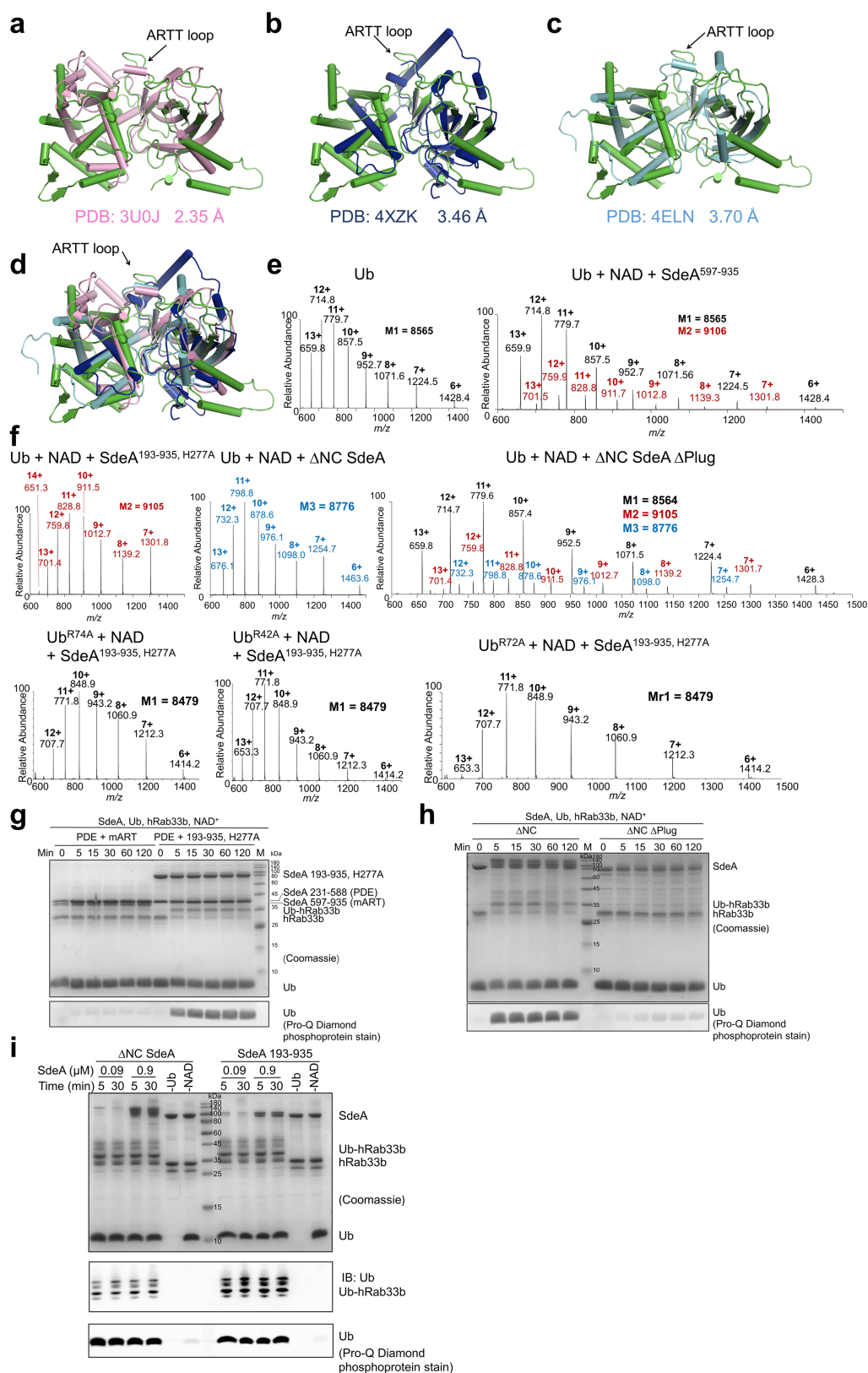
approximately 106 kDa. The buffer is 10 mM Tris pH 8.0, 200 mM NaCl and 5 mM DTT. **e**, Gel filtration profile of the SdeA(231–1190) protein and the molecular markers on Superdex-75 column (GE Healthcare) are shown. The sizes of the molecular markers are marked on top of the peaks. The samples of SdeA(231–1190) collected from the Superdex-75 column were run on SDS–PAGE gels and detected by Coomassie staining. **a–e**, Similar results were obtained in three independent experiments. **a–c, e**, Uncropped blots and gel images are shown in Supplementary Fig. 1. **f**, Two views of the superimposition of the structures of the two molecules in the asymmetric unit, coloured in different colours. **g**, Structure of the CTD region in the crystallized protein can be divided into two parts (left and right). The α helices are numbered according to their orders in the residue region from 908 to 1190. **h**, Topological diagram of the CTD region shown in **g**. The N and C termini of the pCTD domain are labelled.



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Interactions between SdeA mART and SdeA PDE are essential for the activity of SdeA mART. **a**, Overview of the interactions between SdeA mART and SdeA PDE. SdeA is coloured as in Fig. 1b. The major interaction region between the two domains is outlined. **b**, An expanded view of the region outlined in **a**. Interaction residues are shown in stick representation and the red dashed lines represent polar interactions. The plug loop in SdeA mART is indicated. **c**, The interaction between SdeA mART and SdeA PDE. SdeA mART and SdeA PDE are shown in cartoon and surface electrostatic models, respectively. **d**, A view of the interaction from **c** rotated by 180 degrees. In this view, SdeA mART and SdeA PDE are shown as surface electrostatic and cartoon models, respectively. **e**, Testing the ability of SdeA PDE to process ADPR-Ub into

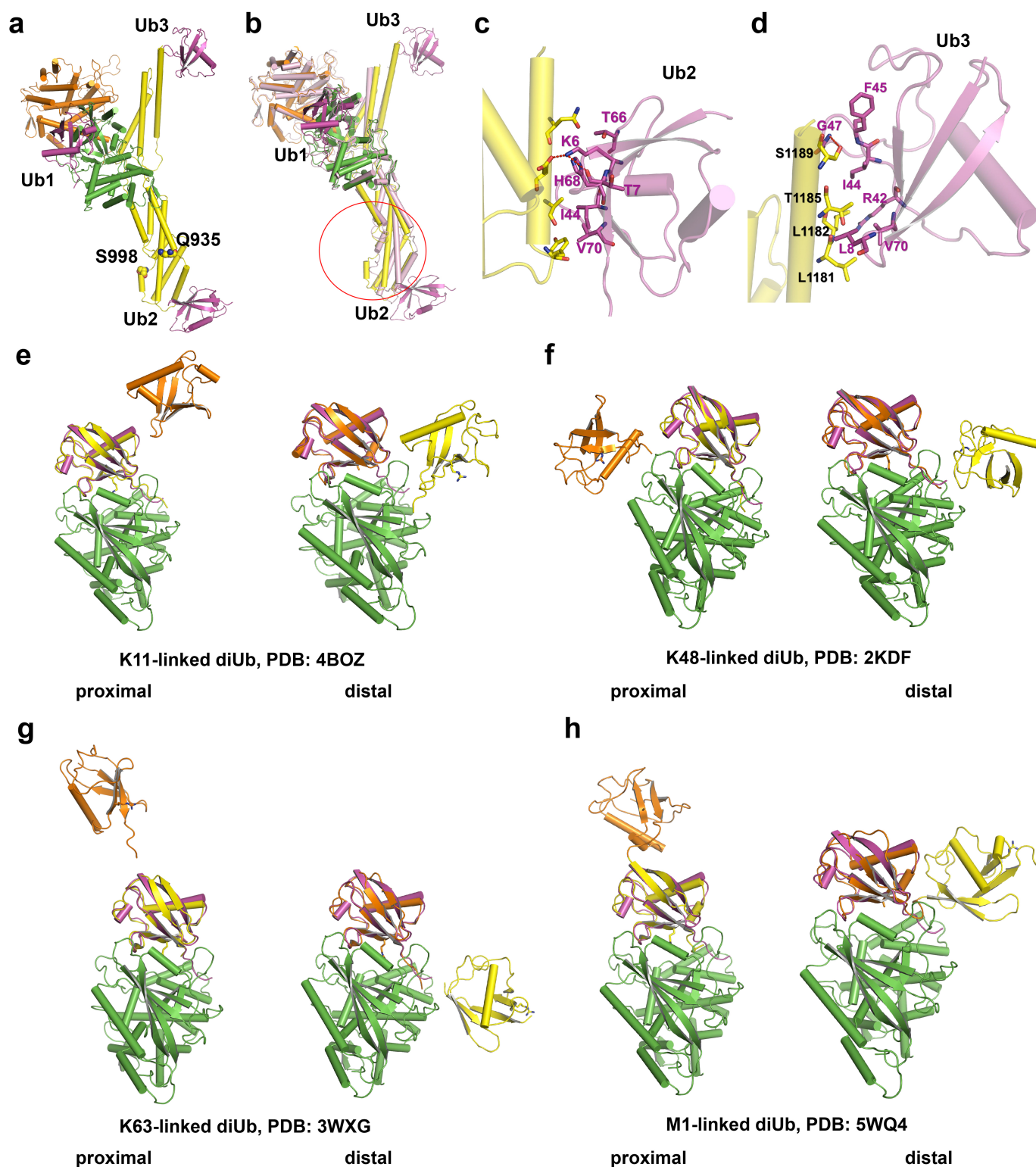
PR-Ub. SdeA(231–588), wild-type Δ NC SdeA or the H277A mutant were incubated with ADPR-Ub and RAB33B for 30 min. The samples were stained with Coomassie and Pro-Q diamond phosphoprotein stain. **f**, Testing the importance of domain interaction for the activity of SdeA mART. Various SdeA segments, and mixtures of SdeA(231–588) and SdeA(597–935) or SdeA(193–935)^{H277A}, were incubated with RAB33B, NAD⁺ and Ub for 30 min. The samples were analysed using Coomassie staining, immunoblotting with anti-Ub antibodies and Pro-Q diamond phosphoprotein staining. **e**, **f**, Similar results were obtained in three independent experiments. Uncropped blots and gel images are shown in Supplementary Fig. 1.



Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | SdeA mART exhibits novel conformations of ARTT and PN loops. **a**, Superimposition of SdeA mART (green) and HopU1 (pink) from *Pseudomonas syringae* (PDB: 3U0J). The ARTT loop is indicated. The r.m.s.d. value is indicated beside the PDB code (panels **b** and **c** are arranged in the same way). **b**, Superimposition of SdeA mART (green) and ADP-ribosyltransferase Vis (blue) (PDB: 4XZK). **c**, Superimposition of SdeA mART (green) and XopAI from *Xanthomonas axonopodis* pv. *citri* (cyan) (PDB: 4ELN). **d**, Superimposition of SdeA mART structure (green) and the three other structures from **a–c**. **e, f**, Mass spectra of the samples in Fig. 2f. The sample name and their molecular masses are indicated in the figures. **g, h**, Different fragments and different

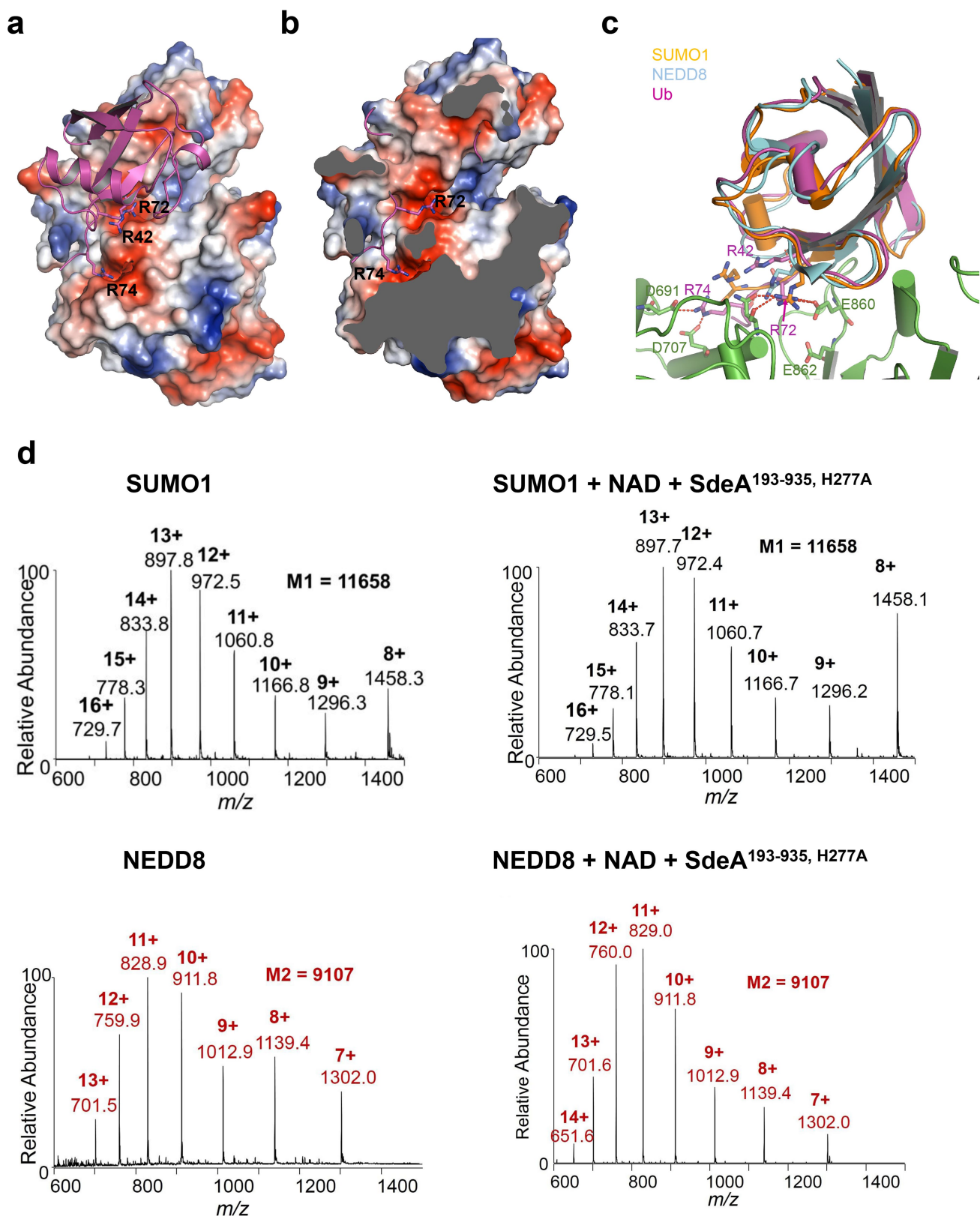
combinations of SdeA proteins were incubated with Ub, RAB33B and NAD^+ at 37 °C for the indicated amounts of time. The samples were analysed using Coomassie staining and Pro-Q diamond phosphoprotein staining. **i**, Testing the ubiquitination ability of the catalytic core. 0.09 or 0.9 μM ΔNC SdeA or SdeA(193–935) was incubated with or without Ub and NAD^+ for the indicated amounts of time. The samples were analysed using Coomassie staining, immunoblotting with anti-Ub antibodies and Pro-Q diamond phosphoprotein staining. **e–i**, Similar results were obtained in three independent experiments. **g–i**, Uncropped blots and gel images are shown in Supplementary Fig. 1.



Extended Data Fig. 5 | SdeA(231-1190) binds three Ub molecules.

a, Overall structure of the SdeA(231-1190)-Ub complex. SdeA is coloured as in Fig. 1b. The three Ub molecules are coloured in magenta and labelled as Ub1-3 according to the order of their binding region in SdeA(231-1190). Q935 and S998, which are two common C termini of the clones used in this study, are shown as spheres. **b**, Ub binding causes prominent structural changes of SdeA. The SdeA-Ub complex structure is shown as in **a**, and the apo-SdeA structure is coloured in pink. The N-terminal

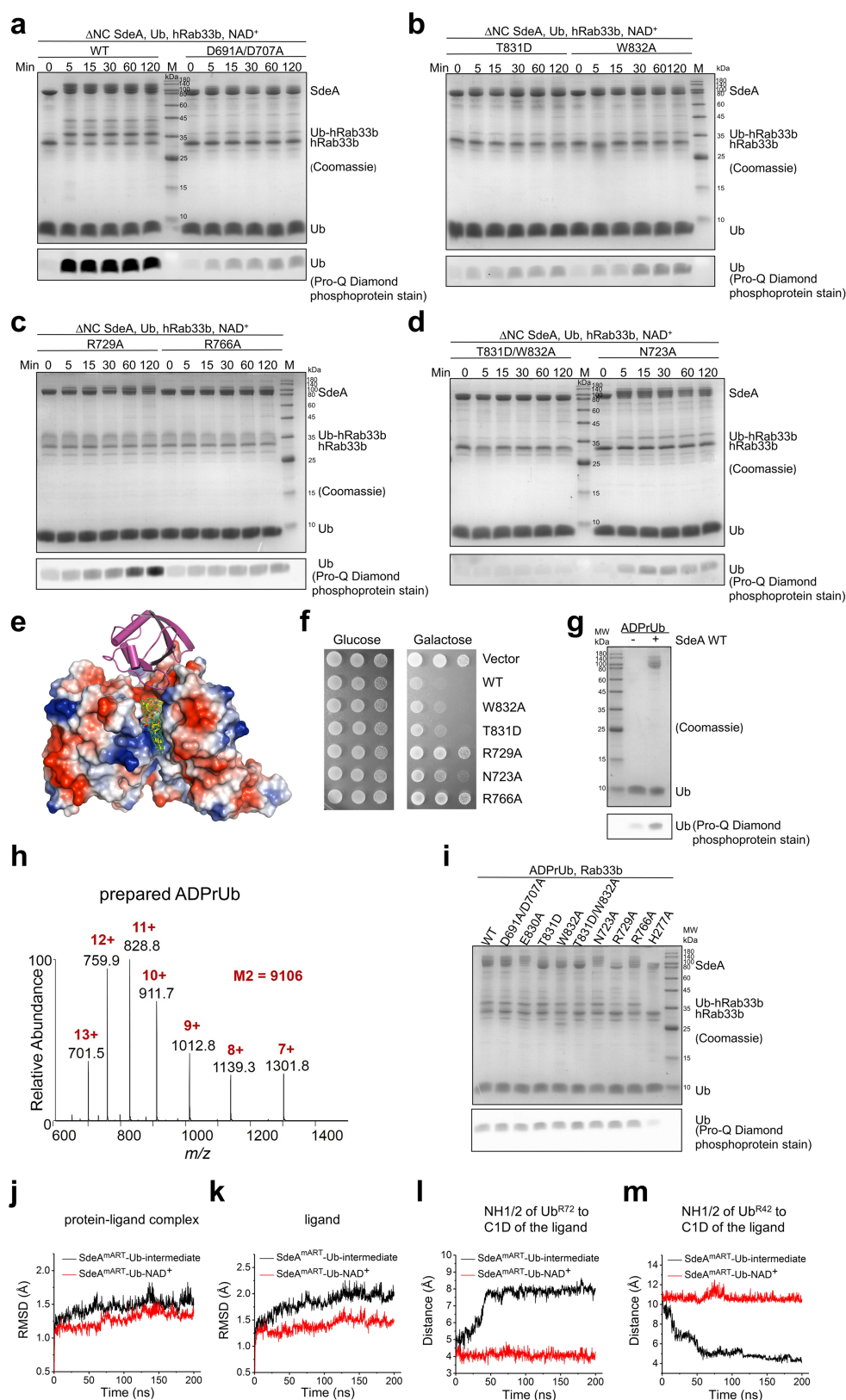
region of SdeA pCTD which undergoes pronounced conformational changes is outlined with a circle. **c**, **d**, Expanded views of the two Ub binding sites in SdeA pCTD. The proteins are coloured as in **a**. Red dashed lines indicate polar interactions. **e-h**, Structural alignments of the Ub molecule (magenta) in the SdeA mART-Ub complex with the proximal (yellow) and distal (orange) Ubs of the K11- (**e**), K48- (**f**), K63- (**g**) and M1-linked (**h**) diubiquitins. The two R42 residues in each of the four diUbs are shown in stick representation.



Extended Data Fig. 6 | Specific recognition of Ub by SdeA mART.

a, The interaction between SdeA mART and Ub. SdeA mART is shown as a surface electrostatic potential model and Ub is in magenta cartoon representation. The R42, R72 and R74 residues of Ub are shown in stick representation. **b**, Ub^{R72} and Ub^{R74} are bound in the negatively charged groove of SdeA mART. The front part of SdeA mART is cut away to reveal the inner surface. **c**, Superimposition of SUMO1 (PDB: 1WM3), NEDD8 (PDB: 1NDD) and Ub in the SdeA mART-Ub complex. The conserved Arg residues in Ub, SUMO1 and NEDD8 are shown in stick representation, out

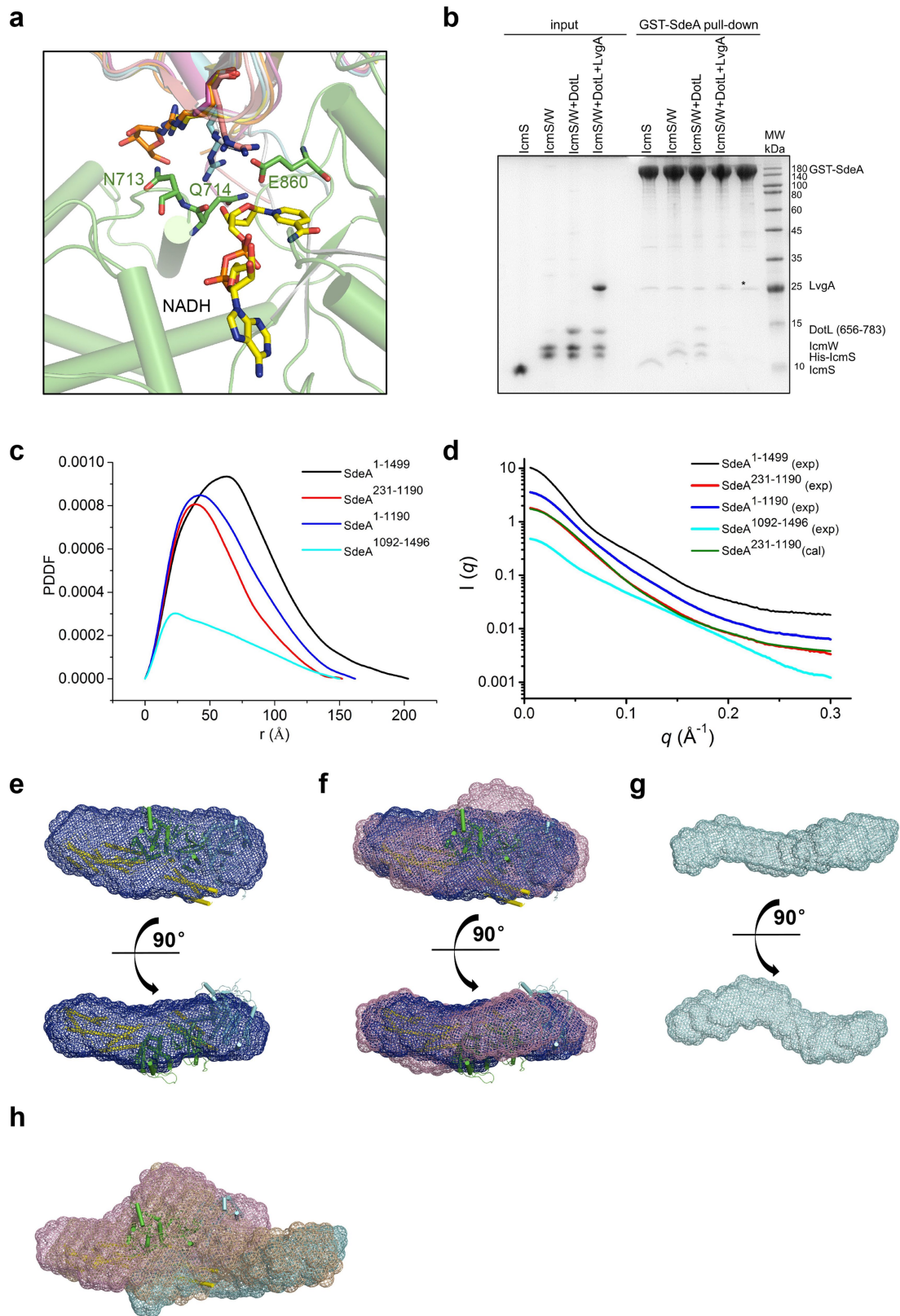
of which Ub^{R42}, Ub^{R72}, and Ub^{R74} are marked. The polar interactions with Ub^{R72} and Ub^{R74} are shown as red dashed lines. **d**, The purified SUMO1 and NEDD8 proteins were incubated with SdeA(193–935)^{H277A} and NAD⁺ under the conditions stated in the ‘Top-down LC–MS analysis of modified Ub and Ub-like proteins’ section of the Methods. Mass spectra of the samples are also shown. The sample names and their molecular masses are indicated in the figures. Similar results were obtained in three independent experiments.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Molecular dynamics simulations indicate the movements of the side chains of Ub^{R42} and Ub^{R72}. **a–d**, Wild-type Δ NC SdeA and indicated mutants were incubated with Ub, RAB33B and NAD⁺ at 37 °C for the indicated amounts of time. The samples were analysed using Coomassie staining and Pro-Q diamond phosphoprotein staining. Similar results were obtained in three independent experiments. Uncropped blots and gel images are shown in Supplementary Fig. 1. **e**, The structure of the SdeA mART–Ub–NADH complex. SdeA mART is shown as an electrostatic surface potential model. White, blue and red indicate neutral, positive and negative surfaces, respectively. Shown in green mesh is the $2F_o - F_c$ electron density map contoured at 1σ around the NADH molecule. **f**, Galactose-inducible pYES2 plasmids containing wild-type Δ NC SdeA or the mutants were transformed into yeast W303 strain. Five microlitres of cells in three tenfold serial dilutions were spotted on both glucose- and galactose-containing plates lacking uracil for two days before image acquisition. **g**, Purified ADPR-Ub proteins were treated with or without wild-type Δ NC SdeA. The samples were analysed using

Coomassie staining and Pro-Q diamond phosphoprotein staining. **h**, Purified ADPR-Ub protein was subjected to top-down LC–MS analysis. The results indicated 100% ADPR-Ub. **i**, Wild-type Δ NC SdeA or other mutants were incubated with RAB33B and the prepared ADPR-Ub verified in **g** and **h**. The samples were analysed using SDS–PAGE, with Coomassie staining and Pro-Q diamond phosphoprotein staining. **f–i**, Similar results were obtained in three independent experiments. **g**, **i**, Uncropped blots and gel images are shown in Supplementary Fig. 1. **j**, **k**, The time series for the r.m.s.d. of the non-hydrogen atoms of the protein–ligand complex (**j**) and the ligand (**k**) in the SdeA mART–Ub–intermediate and SdeA mART–Ub–NAD⁺ systems during molecular dynamics simulations. These two plots indicate that both systems have reached equilibrium during the 200-ns simulations. **l**, **m**, The time series for the shortest distance between the NH1/2 atom of Ub^{R72} and C1D of the ligand (**l**) and the distance between the NH1/2 atom of Ub^{R42} and C1D of the ligand (**m**) in the two systems during molecular dynamics simulations.



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | The overall shape of SdeA and the function of SdeA CTD. **a**, Superimposition of various Ub structures (PDB codes: 5M93 (orange), 1UBQ (pink), 5CRA (chain C, cyan), 3ZLZ (chain B, yellow) and 4BOZ (chain B, grey)) onto the SdeA mART–Ub–NADH structure with R42 residues of all the Ub molecules shown in stick representation. SdeA mART and Ub from the SdeA mART–Ub–NADH complex are shown in green and magenta, respectively. **b**, In vitro GST pull-down assays to detect the interactions of SdeA with IcmS or its complexes. GST-fused SdeA protein was incubated with IcmS, the IcmS–IcmW complex, the IcmS–IcmW–DotLc (residues 656–783 of DotL) ternary complex or the IcmS–IcmW–DotLc–LvgA quaternary complex. The protein samples bound to glutathione resins were washed three times and analysed by SDS–PAGE and Coomassie blue staining. IcmS/W represents IcmS + IcmW. The band marked with an asterisk represents

the degraded GST tag. Similar results were obtained in three independent experiments. Uncropped blots and gel images are shown in Supplementary Fig. 1. **c**, Experimental PDDFs (pair distance distribution function) for SdeA(231–1190), SdeA(1–1499), SdeA(1–1190) and SdeA(1092–1496). **d**, Overlay of the experimental scattering profiles (exp) from the four samples in the SAXS analysis with the back-calculated scattering profile of the crystal structure of SdeA(231–1190) (cal). **e**, Fitting the crystal structure of SdeA(231–1190) into the SAXS envelope of SdeA(231–1190). Two perpendicular views are shown. **f**, Superimposition of the SAXS envelopes of SdeA(231–1190) (coloured as in **e**) and SdeA(1–1190) (light magenta) with the crystal structure of SdeA(231–1190) fitted. **g**, SAXS envelopes of SdeA(1092–1496). **h**, Superimposition of the SAXS envelopes of SdeA(1–1190) (light magenta), SdeA(1092–1496) (cyan) and SdeA(1–1499) (wheat) with the crystal structure of SdeA(231–1190) fitted.

Extended Data Table 1 | Data collection and refinement statistics

	SdeA (5YIM)	SdeA-Ub (5YIK)	SdeA-Ub-NADH (5YIJ)
Data collection			
Space group	C222 ₁	C2	C2
Cell dimensions			
<i>a</i> , <i>b</i> , <i>c</i> (Å)	139.7, 295.4, 194.6	108.9, 145.9, 104.1	107.8, 145.9, 103.4
(°)	90.00, 90.00, 90.00	90.00, 104.46, 90.00	90.00, 103.82, 90.00
Resolution (Å)	50-3.39 (3.52-3.39) ^a	50-3.10 (3.21-3.10)	50-3.18 (3.31-3.18)
<i>R</i> _{sym} or <i>R</i> _{merge} (%)	6.9 (95)	11.1 (76.8)	8.2 (79.1)
<i>I</i> /σ <i>I</i>	14.8 (1.43)	16.0 (2.41)	25.9 (4.2)
Completeness (%)	99.7 (99.7)	99.9 (100.0)	99.7 (100.0)
Redundancy	3.8 (3.9)	5.3 (5.3)	6.5 (6.7)
Refinement			
Resolution (Å)	50.0-3.39 (3.52-3.39)	42.54-3.10 (3.21-3.10)	39.96-3.18 (3.30-3.18)
No. reflections	55636 (5274)	25234 (1380)	24125 (1486)
<i>R</i> _{work} / <i>R</i> _{free}	0.2510/0.2870	0.2230/0.2790	0.2261/0.2752
No. atoms	14128	9381	9410
Protein	14128	9381	9366
Ligand/ion	0	0	44
Water	0	0	0
<i>B</i> factors	132.80	55.0	50.7
Protein	132.80	53.0	50.7
Ligand/ion			45.4
Water			
R.m.s. deviations			
Bond lengths (Å)	0.004	0.008	0.006
Bond angles (°)	0.96	1.02	0.93

For each structure one crystal was used.

^aValues in parentheses are for the highest-resolution shell.

Extended Data Table 2 | Data collection and structural parameters derived from SAXS experiments

	SdeA ¹⁻¹⁴⁹⁹	SdeA ²³¹⁻¹¹⁹⁰	SdeA ¹⁻¹¹⁹⁰	SdeA ¹⁰⁹²⁻¹⁴⁹⁶
<i>Data Collection Parameters</i>				
Facilities and parameters	Settings and values			
Beam line	12ID-B (APS, ANL)		BL19U2 (SSRF)	
Wavelength (Å)	0.8857		1.033	
Detector	Pilatus 1M (SAXS)		Pilatus 100K (SAXS)	
<i>q</i> range (Å ⁻¹)	0.006-2.5		0.009-0.415	
Exposure time (s)	60		60	
Concentration range (mg/ml)	0.75-3		----	
Temperature (K)	300		300	
<i>Structural Parameters</i>				
<i>R</i> _g from guinier fitting (Å)*	54.96±1.66	41.66±1.15	45.40±1.17	43.50±0.35
<i>R</i> _g from GNOM (Å)*	55.83±0.28	42.58±0.27	47.34±0.21	44.97±0.14
<i>D</i> _{max} (Å)	203	148	162	150
<i>V</i> _{porod} from PRIMUS (Å ³)	330.28×10 ³	185.69×10 ³	209.82×10 ³	81.73×10 ³
<i>V</i> _c (Å ²)	1098.56	728.42	818.11	487.69
^a <i>MW</i> ^{Pred} (kDa)	170.1	108.6	133.5	46.4
^b <i>MW</i> ^{saxs1} (kDa)	175.60	101.23	114.85	42.94
^c <i>MW</i> ^{saxs2} (kDa)	206.43	116.06	131.14	51.08
NSD of DAMMIN Models*	0.849±0.025	0.705±0.021	0.763±0.020	0.804±0.019
<i>Software Employed</i>				
Primary Data Processing	Igor Pro/PRIMUS			
<i>P</i> (<i>r</i>) Function	GNOM			
<i>Ab initio</i> Shape Analysis	DAMMIN			
SAXS Profile Computation	CRY SOL			
Superposition and averaging	SUPCOMB/DAM AVER			

*The predicted molecular mass (*MW^{Pred}*) was calculated from the primary sequences of components.

^bThe molecular mass from SAXS1 (*MW^{saxs1}*) was calculated using the previously developed *R_g*/*V_c* power law⁴³.

^cThe molecular mass from SAXS2 (*MW^{saxs2}*) was calculated from Porod volume using AUTOPOROD⁶⁰.

*Data are mean ± s.d. from fitting.

Spatiotemporal regulation of liquid-like condensates in epigenetic inheritance

Gang Wan^{1,4}, Brandon D. Fields^{1,2,4}, George Spracklin^{1,2}, Aditi Shukla¹, Carolyn M. Phillips³ & Scott Kennedy^{1*}

Non-membrane-bound organelles such as nucleoli, processing bodies, Cajal bodies and germ granules form by the spontaneous self-assembly of specific proteins and RNAs. How these biomolecular condensates form and interact is poorly understood. Here we identify two proteins, ZNFX-1 and WAGO-4, that localize to *Caenorhabditis elegans* germ granules (P granules) in early germline blastomeres. Later in germline development, ZNFX-1 and WAGO-4 separate from P granules to define an independent liquid-like condensate that we term the Z granule. In adult germ cells, Z granules assemble into ordered tri-condensate assemblages with P granules and *Mutator* foci, which we term PZM granules. Finally, we show that one biological function of ZNFX-1 and WAGO-4 is to interact with silencing RNAs in the *C. elegans* germline to direct transgenerational epigenetic inheritance. We speculate that the temporal and spatial ordering of liquid droplet organelles may help cells to organize and coordinate the complex RNA processing pathways that underlie gene-regulatory systems, such as RNA-directed transgenerational epigenetic inheritance.

Epigenetic information can be inherited for several generations (transgenerational epigenetic inheritance, TEI)^{1,2}. Non-coding RNAs have emerged as important mediators of TEI (RNA-directed TEI), although the mechanism(s) by which RNA mediates TEI remains poorly understood. In many eukaryotes, double-stranded RNAs (dsRNAs) silence other cellular RNAs that exhibited sequence complementarity to trigger dsRNAs; a process termed RNA interference (RNAi)³. In *C. elegans*, RNAi is heritable: distant progeny of animals exposed to dsRNAs continue to silence complementary RNAs in the absence of further dsRNA exposure (RNAi inheritance)^{4–6}. To further our understanding of RNA-directed TEI, we conducted a genetic screen to identify factors required for RNA inheritance (Extended Data Fig. 1). Our screen identified 37 mutations that disrupted RNAi inheritance. We subjected DNA from these 37 mutant strains to whole-genome sequencing and identified four independent mutations in the gene *zk1067.2* (Fig. 1a). To confirm that *zk1067.2* is required for RNAi inheritance, we tested two additional alleles of *zk1067.2* (*gk458570* and *gg561*) for defects in RNAi inheritance. *gk458570* and *gg561* animals responded normally to dsRNA treatment; however, the progeny of these mutant animals were largely unable to inherit gene silencing (Fig. 1b and Extended Data Fig. 2). We conclude that *zk1067.2* is required for RNAi inheritance.

ZNFX-1 is required for TEI

Sequence analysis showed that ZK1067.2 is predicted to encode a 2,443-amino acid protein that contains a superfamily one (SF1) RNA helicase domain and a zinc-finger domain (Fig. 1a). A single putative orthologue of ZK1067.2 was found in most eukaryotic genomes. Fungal orthologues have been linked to RNAi pathways in *Schizosaccharomyces pombe* and *Neurospora crassa*^{20,26}. Homology between ZK1067.2 and its mammalian orthologue ZNFX1 extend to a zinc-finger domain not present in fungal orthologues. We conclude that ZK1067.2 is a conserved protein involved in RNAi-mediated gene silencing in many eukaryotes. Hereafter, we refer to ZK1067.2 as ZNFX-1.

To begin to understand the function of ZNFX-1 during RNAi inheritance, we used CRISPR-Cas9 to insert a *3xflag::gfp* epitope immediately upstream of the *znfx-1* start codon. Note, CRISPR-mediated

gene conversion was used throughout this work. Tagged loci were expressed near wild-type levels and resultant fusion proteins were functional unless otherwise indicated (Extended Data Fig. 3). We observed GFP::ZNFX-1 expression in the adult germline as well as in developing germ cells during all stages of embryonic and larval development (Fig. 1c). No GFP::ZNFX-1 expression was observed in somatic tissues. After fertilization, *C. elegans* zygotes undergo a series of asymmetric cell divisions in which germline determinants segregate with germline blastomeres. During embryonic development, ZNFX-1 foci were concentrated in, and segregated with, the germline blastomeres (Fig. 1c and see below). In adult germ cells, GFP::ZNFX-1 was concentrated in foci that were distributed in a perinuclear pattern around nuclei (Fig. 1d). We conclude that *znfx-1* encodes a germline-expressed protein that segregates with the germline and localizes to perinuclear foci in adult germ cells.

Treatment of animals with *oma-1* dsRNA silences the *oma-1* gene for several generations^{5,6}. To address when ZNFX-1 acts to promote RNAi inheritance, we used quantitative reverse transcription PCR (qRT-PCR) to measure *oma-1* mRNA and precursor mRNA (pre-mRNA) levels in *znfx-1*(–) animals exposed to *oma-1* RNAi, as well as in the progeny of these animals. *znfx-1*(–) animals responded normally to *oma-1* RNAi; however, their progeny failed to inherit silencing, suggesting that ZNFX-1 acts during the inheritance phase of RNAi (Fig. 1e). During RNAi inheritance, short interfering RNAs (siRNAs) that target genes undergoing RNAi silencing are expressed for several generations⁶. In *znfx-1*(–) animals exposed directly to *oma-1* dsRNA, *oma-1* siRNAs were produced at wild-type levels; however, the progeny of these mutant animals failed to express *oma-1* siRNAs (Fig. 1f). Three additional genetic and biochemical analyses supported the idea that ZNFX-1 acts during the inheritance phase of RNAi (Extended Data Fig. 4). These data establish that ZNFX-1 is a dedicated RNAi inheritance factor.

WAGO-4 acts with ZNFX-1 to direct TEI

The *C. elegans* genome encodes approximately 27 Argonaute (AGO) proteins. The molecular function of many of these AGOs remains

¹Department of Genetics, Harvard Medical School, Boston, MA, USA. ²Laboratory of Genetics, University of Wisconsin-Madison, Madison, WI, USA. ³Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA. ⁴These authors contributed equally: Gang Wan, Brandon D. Fields. *e-mail: kennedy@genetics.med.harvard.edu

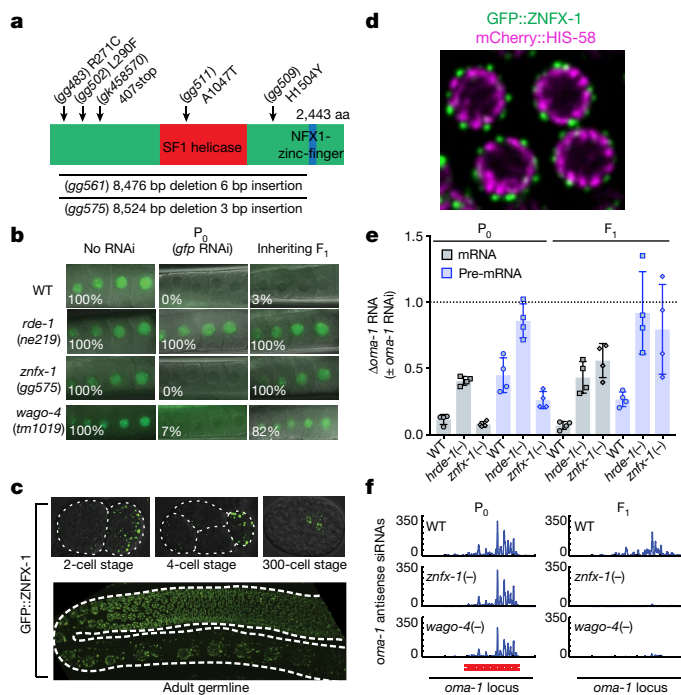


Fig. 1 | ZNF-1 is a conserved RNA helicase required for RNAi inheritance in *C. elegans*. **a**, *znfx-1* alleles are indicated. aa, amino acids. **b**, Animals expressing a *pie-1::gfp::h2b* transgene¹⁷ were exposed to *gfp* dsRNA. F₁ progeny were grown in the absence of dsRNA, and GFP expression in oocytes was visualized by fluorescence microscopy using a 63× objective. The percentage of animals that express *pie-1::gfp::h2b* is shown (*n* = 6 biologically independent samples for wild type (WT), *n* = 3 for *rde-1*, *znfx-1* and *wago-4*, each *n* ≥ 30 animals). P₀, parental generation. **c**, Fluorescence micrograph of *gfp::znfx-1* animals. Images are representative of more than three animals visualized at each life stage using a 60× objective. **d**, Pachytene germ cells of animals that express GFP::ZNF-1 and the chromatin marker mCherry::HIS-58. Image is representative of three animals, visualized using a 60× objective. **e**, Wild-type, *hrde-1*(*tm1200*) and *znfx-1*(*gg561*) animals were exposed to *oma-1* dsRNA. Total RNA from RNAi (P₀) and inheriting (F₁) generations was isolated. RNA was quantified by qRT-PCR using primers 5' to the site of RNAi and data were normalized to *eft-3* pre-mRNA. *n* = 4 biologically independent samples. Data are mean ± s.d. Note that independent mRNA and pre-mRNA primer sets gave similar results (data not shown). **f**, siRNA libraries (see Methods) were prepared from wild-type, *znfx-1*(*gg561*) and *wago-4*(*tm1019*) animals exposed to *oma-1* dsRNA (P₀) and progeny (F₁). Antisense reads mapping to *oma-1* locus are shown. Red line indicates region of *oma-1* locus targeted by dsRNA. Reads counts were normalized to total number of sequenced reads (*n* = 2 biologically independent samples).

unknown. Two of the mutant strains identified by our genetic screen harboured mutations in the AGO-encoding gene *wago-4*. To confirm WAGO-4 is required for RNAi inheritance, we tested two additional *wago-4* deletion alleles (*tm1019* and *tm2401*) for RNAi inheritance defects. Both alleles exhibited RNAi inheritance defects (Fig. 1b and Extended Data Fig. 5). Thus, like ZNF-1, WAGO-4 is required for RNAi inheritance. Furthermore, when we appended a *gfp* tag to the *wago-4* locus, we observed that, like ZNF-1, WAGO-4 is a germline-expressed protein that segregates with the P lineage blastomeres and localizes to perinuclear foci (Extended Data Fig. 5). For unknown reasons, our GFP::WAGO-4 fusion protein was fully functional for RNAi inheritance in some RNAi inheritance assays but only partially functional in other assays (Extended Data Fig. 3). TagRFP::ZNF-1 and GFP::WAGO-4 colocalized in germ cells, suggesting that WAGO-4 and ZNF-1 may act together to promote RNAi inheritance (Fig. 2a). Three additional lines of evidence support this idea. First, Flag-tagged WAGO-4 (3×Flag::WAGO-4) co-precipitates with haemagglutinin-tagged ZNF-1 (HA::ZNF-1), but not with

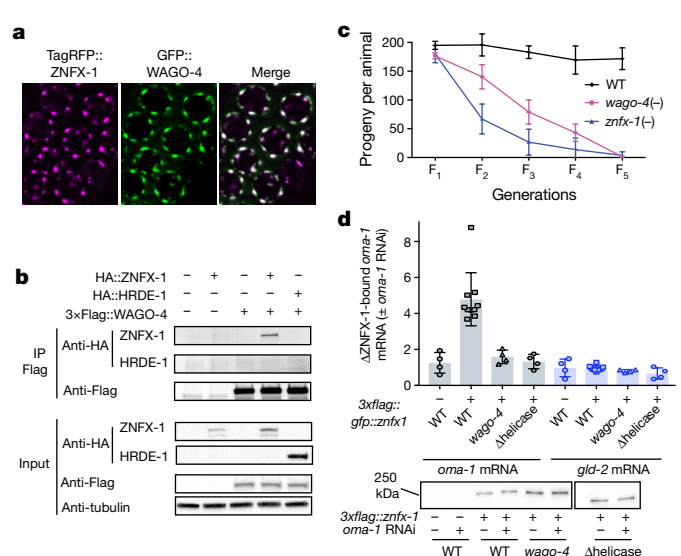


Fig. 2 | ZNF-1 and WAGO-4 act cooperatively to drive RNAi inheritance. **a**, Fluorescent micrographs of pachytene germ cells that express GFP::WAGO-4 and TagRFP::ZNF-1. Image is representative of more than three animals. **b**, Co-immunoprecipitation analysis of HA::ZNF-1 and Flag::WAGO-4. HA::HRDE-1 is a negative control. *n* = 1 for wild-type and HA::HRDE-1, *n* = 3 for others; *n* denotes independent experiments. **c**, Animals of indicated genotypes were shifted to growth at 25°C and progeny were counted for five generations. Data are mean ± s.d. of *n* = 3 biologically independent samples. **d**, Top, Flag::ZNF-1 was immunoprecipitated in RNAi generation from animals treated with or without *oma-1* dsRNA. Co-precipitating RNA was subjected to qRT-PCR to quantify *oma-1* mRNA co-precipitating with ZNF-1 in wild-type or *wago-4*(*tm1019*) animals. Note that ZNF-1 also binds RNAi-targeted RNAs in inheriting generations (data not shown). ZNF-1 Δhelicase contains a 1,487 base-pair (bp) in-frame deletion of the ZNF-1 helicase domain. *gld-2* is a control mRNA. *n* = 10 for wild type, *n* = 4 for others; *n* denotes biologically independent samples. Data are mean ± s.d. Bottom, western blot of immunoprecipitated ZNF-1 from one RNAi precipitate replicate shown in the top panel. Two unrelated lanes were removed from this image (see Supplementary Fig. 1).

a haemagglutinin-tagged negative control protein (HA::HRDE-1), suggesting a physical interaction between the two proteins (Fig. 2b). Second, *wago-4* mutant animals behaved like *znfx-1* mutant animals in molecular assays of RNAi inheritance (Fig. 1f). Third, *znfx-1* and *wago-4* animals share a pleiotropic phenotype: both mutant animals exhibited a temperature-sensitive mortal germline (Mrt) phenotype, in which mutant animals became sterile several generations after populations were shifted to growth at a higher temperature (25°C) (Fig. 2c). Taken together, these data show that WAGO-4 functions with ZNF-1 to transmit RNA-based epigenetic information across generations.

How WAGO-4 and ZNF-1 promote RNAi inheritance is unclear. The closest homologue of ZNF-1 is SMG-2 (also known as UPF1), which marks mRNAs containing premature termination codons⁷. We wondered whether, by analogy, ZNF-1 might bind and mark mRNAs encoded by genes undergoing heritable gene silencing. To test this idea, we subjected animals expressing 3×Flag::ZNF-1 to *oma-1* RNAi, immunoprecipitated 3×Flag::ZNF-1, and used qRT-PCR to determine whether *oma-1* RNAi caused ZNF-1 to interact with *oma-1* mRNA. Indeed, *oma-1* RNAi caused ZNF-1 to co-precipitate with *oma-1* mRNA (Fig. 2d). The following three lines of evidence show that the interaction of ZNF-1 with TEI-related RNAs is a sequence-specific event directed by the RNAi machinery. First, RNAi that targets the *lin-15b* gene caused ZNF-1 to interact with the *lin-15b* mRNA, but not the *oma-1* mRNA (and vice versa), indicating that ZNF-1 and mRNA interactions are sequence-specific (data not shown). Second, most RNA helicases bind RNA via their helicase domains. Deletion of the ZNF-1 helicase domain did not affect ZNF-1 expression but did prevent ZNF-1 from interacting with *oma-1* mRNA (Fig. 2d). Third,

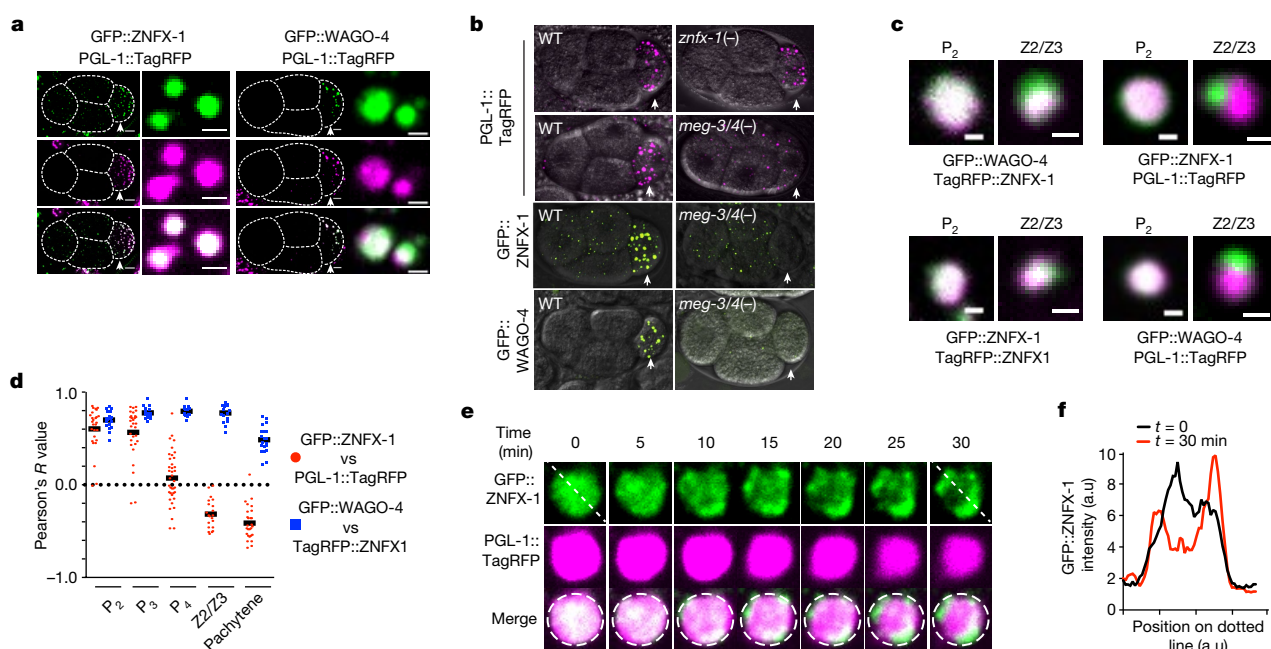


Fig. 3 | ZNFX-1 and WAGO-4 separate from P granules to form new foci during germline development. **a, b,** Micrographs of four-cell embryos expressing indicated proteins are shown. P₂ blastomeres are indicated by arrowheads. **a,** Magnifications of P granules are shown to the right. Images are representative of more than three animals. **b,** Genotypes are *znfx-1(gg561)*, or *meg-3(tm4259); meg-4(ax2026)*. Images are representative of more than three animals. **c, d,** Micrographs (**c**) and quantification (**d**) of colocalization between (see Methods) indicated

fluorescent proteins at indicated stages of germline development. **c,** Images are representative more than three animals. **d,** Each data point represents the quantification of an individual granule ($n > 15$ granules imaged in > 3 animals for each lifestage). Mean is indicated by solid black line. **e,** Time-lapse micrographs in early (approximately 300 cell embryos) Z2/Z3 cells. Images are representative of three animals. **f,** Fluorescent intensity along dotted lines shown in **e**. Scale bars, 6 μ m (**a**, whole embryo), 1 μ m, (**a**, individual granules), and 0.5 μ m (**c**). a.u., arbitrary units.

in *wago-4* mutant animals, ZNFX-1 failed to interact with the *oma-1* mRNA in response to *oma-1* RNAi (Fig. 2d). We conclude that RNAi directs ZNFX-1 to interact with mRNAs undergoing heritable silencing and that WAGO-4 is required for this property of ZNFX-1.

ZNFX-1 and WAGO-4 separate from P granules

P granules are biomolecular condensates that, like ZNFX-1 and WAGO-4 foci, segregate with the germline blastomeres (P₀–P₄) during embryonic development^{8,9}. The low-complexity protein PGL-1 marks P granules¹⁰. GFP::ZNFX-1 and GFP::WAGO-4 colocalized with PGL-1::TagRFP in P₁–P₃ germline blastomeres, suggesting that ZNFX-1 and WAGO-4 are P granule factors (Fig. 3a). MEG-3 and MEG-4 are low-complexity domain proteins that are redundantly required for P granule formation in the P lineage¹¹ (Fig. 3b). In *meg-3/4(-)* embryos, ZNFX-1 and WAGO-4 foci failed to segregate with the P lineage (Fig. 3b). Thus, in early P₁–P₃ germline blastomeres, ZNFX-1 and WAGO-4 localize to P granules.

At around the 100-cell stage of embryonic development, the P₄ blastomere divides to give rise to Z2 and Z3, which are the primordial germ cells of *C. elegans*. Notably, we found that GFP::ZNFX-1 no longer colocalized with PGL-1::TagRFP in Z2 and Z3 (Fig. 3c). Instead, GFP::ZNFX-1 appeared in foci that were adjacent to (see below), yet distinct from, PGL-1::TagRFP foci (Fig. 3c). Similar results were seen when antibodies were used to visualize PGL-1 and ZNFX-1, indicating that failure to colocalize was not an artefact of GFP or TagRFP epitopes (Extended Data Fig. 6). Quantitative analyses showed that the degree to which ZNFX-1 and PGL-1 colocalized changed during development, with a transition from colocalized to non-colocalized occurring between the P₃ and Z2/Z3 cells (Fig. 3d). The ZNFX-1 and WAGO-4 foci seen in Z2 and Z3 could form de novo or by the separation of ZNFX-1, WAGO-4 and PGL-1 from within pre-existing foci. We favour the latter model, as time-lapse imaging in Z2 or Z3 cells captured what appeared to be ZNFX-1 and PGL-1 separation events (Fig. 3e, f). The separation of ZNFX-1 and WAGO-4 into discrete foci could be triggered by phase separation or by segregation of pre-existing

sub-structures into discrete areas¹¹. We conclude that, late in germline development, ZNFX-1 and WAGO-4 are concentrated in foci adjacent to P granules.

Z granules are liquid-like condensates

Liquid-like condensates are self-assembling cellular structures that form when specific proteins and RNAs undergo liquid–liquid phase transitions from surrounding cytoplasm. The ability of ZNFX-1 and WAGO-4 to separate from P granules suggests that ZNFX-1 and WAGO-4 foci may also be liquid-like condensates. Liquid-like condensates are typically spherical in shape and their internal constituents undergo rapid internal rearrangements^{12,13}. Consistent with the idea that ZNFX-1 foci are liquid-like condensates, we observed that during oocyte maturation, ZNFX-1 foci detached from nuclei and assumed spherical shapes (Extended Data Fig. 7). In addition, fluorescence recovery after photobleaching (FRAP) experiments showed that within ZNFX-1 foci, GFP::ZNFX-1 fluorescence recovered rapidly from bleaching ($t = 8$ s), which is a rate similar to that reported for PGL-1 FRAP in P granules⁸ (Extended Data Fig. 7). Thus, ZNFX-1 and WAGO-4 foci (post Z2 or Z3) exhibit properties reminiscent of liquid-like condensates and, therefore, we refer to these foci as Z granules.

Z granules assemble into tri-droplet structures

C. elegans germ cells possess at least two other foci (processing bodies and *Mutator* foci) with properties similar to liquid-like condensates^{14,15}. TagRFP::ZNFX-1 did not co-localize with MUT-16::GFP, which marks *Mutator* foci, nor did GFP::ZNFX-1 colocalize with mCherry::PATR-1 or mRuby::DCAP-1, which mark processing bodies (Fig. 4a and Extended Data Fig. 7). Interestingly, although Z granules did not colocalize with *Mutator* foci, the relative positions of these two foci were not random. Z granules were usually (89% of the time, $n = 35$) found closely apposed to (no empty space between fluorescence signals) a *Mutator* foci (Fig. 4a). Similarly, Z granules were usually (91% of the time, $n = 35$) found closely apposed to a P granule (Fig. 4a). Quantification of distances between surfaces and centres of fluorescence for the three

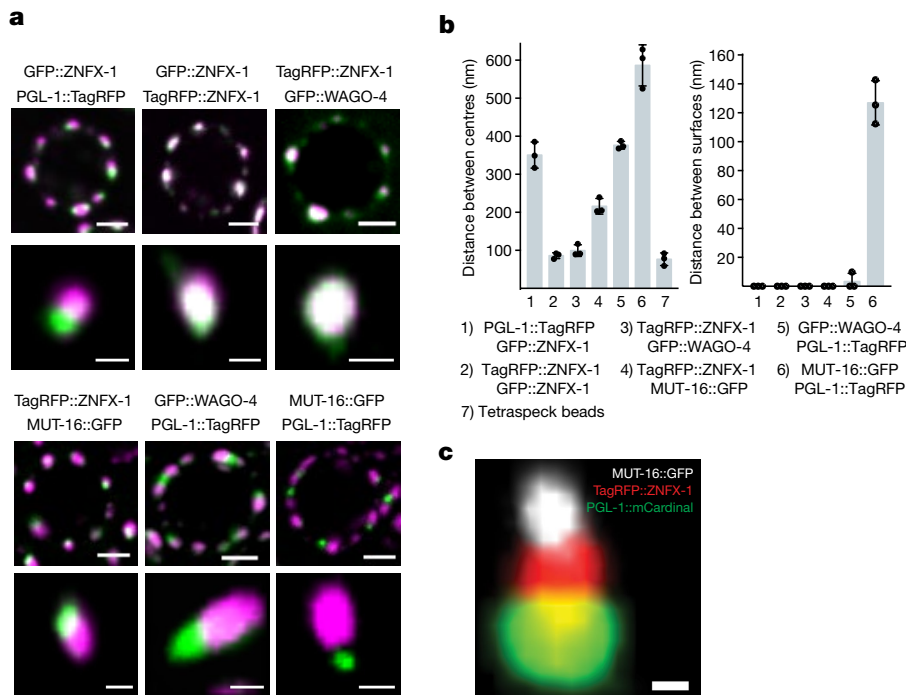


Fig. 4 | Z granules assemble into tri-condensate (PZM) structures with P granules and *Mutator* foci. **a**, Top, fluorescent micrographs of a pachytene germ-cell nucleus from animals expressing the indicated fluorescent proteins. Bottom, 3D renders of representative foci. Images are representative of more than three animals. **b**, Distance between the centres (left) and surfaces (right) of the spaces occupied by the indicated fluorescent proteins was calculated as described in the Methods. Data are mean \pm s.d. of 10 granules measured in 3 animals (30 total). Column

foci supported the idea that Z granules localize adjacent to P granules and *Mutator* foci in adult germ cells (Fig. 4b). This analysis also showed that the distance between the surfaces of Z granules and P granules or *Mutator* foci (but not the distance between P granules and *Mutator* foci) lies within the diffraction limit of light, indicating that Z granules exist in very close proximity to, and may be in direct physical contact with P granules and *Mutator* foci (Fig. 4b). Note that although Z granules are intimately associated with P granules and *Mutator* foci in adult germ cells (and throughout most of germline development), they can exist independently. For instance, in the adult germline, Z granules remained visible at developmental time points when P granules were no longer present (Extended Data Fig. 7). Similarly, Z granules are present in developing germ cells at time points (that is, P blastomeres) when *Mutator* foci are not thought to be present¹⁴. In addition, shearing force causes P granules in pachytene-stage germ cells to disengage from nuclei and flow through the germline syncytium⁸. After applying shearing force, we found that P granules flowed through the cytoplasm; however, Z granules remained largely static (Extended Data Fig. 7). Thus, Z granules can be separated from P granules and *Mutator* foci both developmentally and physically. We conclude that Z granules represent an independent form of liquid-like condensate, which closely mirror P granules and *Mutator* foci in adult germ cells.

Our data suggest that Z granules may localize between (bridge) P granules and *Mutator* foci. To test this idea, we imaged the three foci simultaneously using animals that express PGL-1::mCardinal, TagRFP::ZNFX-1 or MUT-16::GFP¹⁶. This analysis confirmed the idea that Z granules bridge P granules and *Mutator* foci (Fig. 4c). In 60% (52 out of 86) of cases, we observed a Z granule in close apposition to both a P granule and a *Mutator* foci, whereas in 92% (48 out of 52) of these cases, the Z granule lay between the other two foci. In no case (0 out of 52) did a P granule or a *Mutator* foci bridge the other two types of foci, respectively. Quantification of the distances between the centres and surfaces of Z granules, P granules and *Mutator* foci from

7 shows a chromatic shift associated with tetraspeck beads. Data in the right panel have been corrected for this shift. **c**, Fluorescent micrograph of indicated fluorescent proteins from pachytene germ cell. Image is representative of more than three animals (see Extended Data Fig. 8 for quantifications). Scale bars, 2 μ m (**a**, germ cells), 0.5 μ m (**a**, single granules), and 0.25 μ m (**c**). Positions of nuclear membrane, nuclear pores and PZM segments are not yet known.

triple-marked images support the idea that Z granules act as a bridge between P granules and *Mutator* foci in adult germ cells (Extended Data Fig. 8). We conclude that P granules, Z granules and *Mutator* foci form tri-condensate assemblages (henceforth termed PZMs) in adult germ cells, and that the relative position of the three liquid-like condensates constituting the PZM is ordered.

Following these observations, we sought to determine whether PZM assembly was required for RNA-directed TEI. Factors concentrated in *Mutator* foci^{14,17–19} and Z granules (this work) contribute to RNA-directed TEI. Furthermore, we find that several factors, known to be required for P granule assembly, are also needed for efficient RNAi inheritance (Extended Data Fig. 9). Thus, factors associated with all three segments of the PZM have now been linked to TEI. We also find that in mutant animals with defective P granules, Z granules become malformed and ZNFX-1 fails to bind TEI-related RNAs, hinting that the segments of the PZM may communicate with each other during TEI (Extended Data Fig. 9). The results are consistent with the idea that PZM assembly is important for RNA-directed TEI.

Discussion

We show here that the inheritance factors ZNFX-1 and WAGO-4 localize to a liquid-like condensate that we name the Z granule. Given that Z granules segregate with the germline, we speculate that one function of the Z granule is to concentrate and segregate silencing factors into the germline to promote RNA-based TEI. ZNFX-1 is a conserved RNA helicase that localizes to Z granules and marks RNAs produced from genes undergoing heritable silencing. The *S. pombe* orthologue of ZNFX-1 is Hrr1, which forms a nuclear complex (termed the RDRC) with Argonaute and RdRP to amplify siRNA populations directing pericentromeric heterochromatin²⁰. We speculate that a *C. elegans* version of the RDRC acts in the cytoplasm where it promotes RNAi inheritance by: (1) binding inherited siRNAs (via WAGO-4), (2) marking mRNAs complementary to inherited siRNAs (via ZNFX-1), (3) using marked

mRNAs as templates for RdRP-based siRNA amplification, and (4) repeating this cycle each generation (Extended Data Fig. 10). Note, a related study suggests that the function of ZNFX-1 in RNA marking may involve positioning RdRP enzymes to prevent 5' drift of AGOs targeting mRNAs²¹.

ZNFX-1 and WAGO-4 separate from components of the P granule during early embryogenesis to form an independent liquid-like condensate. Separation occurs at a developmental time that roughly correlates with the first association of P granules with nuclear pores and the advent of germline transcription^{9,22–25}. We speculate that condensate separation might be triggered when newly synthesized mRNAs transit P granules and interact with RNA-binding proteins to alter local protein concentration and initiate separation. In addition to temporal ordering, we find that Z granules are spatially ordered relative to P granules and *Mutator* foci, with Z granules forming the centrepiece of PZM tri-condensate assemblages in adult germ cells. These results show that mechanism(s) exist to organize and arrange liquid-like condensates in space as well as time. Further work is needed to understand how PZM segments assemble in the correct order and to determine whether/how PZM assembly contributes to RNA-based TEI. Small RNA-based pathways in animals are complex with many thousands of small regulatory RNAs that regulate thousands of mRNAs at almost all levels of gene expression. We speculate that the ordering of liquid-like condensates in space and time helps to organize and coordinate these small RNA pathways, including RNA-directed TEI (Extended Data Fig. 10). Similar strategies may be used by cells to organize and coordinate other gene regulatory or biochemical pathways.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0132-0>.

Received: 30 October 2017; Accepted: 10 April 2018;

Published online: 16 May 2018

1. Heard, E. & Martienssen, R. A. Transgenerational epigenetic inheritance: myths and mechanisms. *Cell* **157**, 95–109 (2014).
2. Lim, J. P. & Brunet, A. Bridging the transgenerational gap with epigenetic memory. *Trends Genet.* **29**, 176–186 (2013).
3. Fire, A. et al. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806–811 (1998).
4. Vastenhouw, N. L. et al. Gene expression: long-term gene silencing by RNAi. *Nature* **442**, 882 (2006).
5. Alcazar, R. M., Lin, R. & Fire, A. Z. Transmission dynamics of heritable silencing induced by double-stranded RNA in *Caenorhabditis elegans*. *Genetics* **180**, 1275–1288 (2008).
6. Buckley, B. A. et al. A nuclear Argonaute promotes multigenerational epigenetic inheritance and germline immortality. *Nature* **489**, 447–451 (2012).
7. Brogna, S., McLeod, T. & Petric, M. The Meaning of NMD: translate or perish. *Trends Genet.* **32**, 395–407 (2016).
8. Brangwynne, C. P. et al. Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science* **324**, 1729–1732 (2009).
9. Strome, S. & Wood, W. B. Generation of asymmetry and segregation of germ-line granules in early *C. elegans* embryos. *Cell* **35**, 15–25 (1983).
10. Strome, S. & Wood, W. B. Immunofluorescence visualization of germ-line-specific cytoplasmic granules in embryos, larvae, and adults of *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **79**, 1558–1562 (1982).
11. Wang, J. T. et al. Regulation of RNA granule dynamics by phosphorylation of serine-rich, intrinsically disordered proteins in *C. elegans*. *eLife* **3**, e04591 (2014).
12. Toretzky, J. A. & Wright, P. E. Assemblages: functional units formed by cellular phase separation. *J. Cell Biol.* **206**, 579–588 (2014).
13. Weber, S. C. & Brangwynne, C. P. Getting RNA and protein in phase. *Cell* **149**, 1188–1191 (2012).
14. Phillips, C. M., Montgomery, T. A., Breen, P. C. & Ruvkun, G. MUT-16 promotes formation of perinuclear mutator foci required for RNA silencing in the *C. elegans* germline. *Genes Dev.* **26**, 1433–1444 (2012).
15. Gallo, C. M., Munro, E., Rasoloson, D., Merritt, C. & Seydoux, G. Processing bodies and germ granules are distinct RNA granules that interact in *C. elegans* embryos. *Dev. Biol.* **323**, 76–87 (2008).
16. Chu, J. et al. Non-invasive intravital imaging of cellular differentiation with a bright red-excitable fluorescent protein. *Nat. Methods* **11**, 572–578 (2014).
17. Ashe, A. et al. piRNAs can trigger a multigenerational epigenetic memory in the germline of *C. elegans*. *Cell* **150**, 88–99 (2012).
18. Grishok, A., Tabara, H. & Mello, C. C. Genetic requirements for inheritance of RNAi in *C. elegans*. *Science* **287**, 2494–2497 (2000).
19. Shirayama, M. et al. piRNAs initiate an epigenetic memory of nonself RNA in the *C. elegans* germline. *Cell* **150**, 65–77 (2012).
20. Motamedi, M. R. et al. Two RNAi complexes, RITS and RDRC, physically interact and localize to noncoding centromeric RNAs. *Cell* **119**, 789–802 (2004).
21. Mello, C. et al. ZNFX-1 functions within perinuclear nuage to balance epigenetic signals. *Mol. Cell*. <https://doi.org/10.1016/j.molcel.2018.04.009> (2018).
22. Seydoux, G. & Dunn, M. A. Transcriptionally repressed germ cells lack a subpopulation of phosphorylated RNA polymerase II in early embryos of *Caenorhabditis elegans* and *Drosophila melanogaster*. *Development* **124**, 2191–2201 (1997).
23. Pitt, J. N., Schisa, J. A. & Priess, J. R. P granules in the germ cells of *Caenorhabditis elegans* adults are associated with clusters of nuclear pores and contain RNA. *Dev. Biol.* **219**, 315–333 (2000).
24. Furuhashi, H. et al. Trans-generational epigenetic regulation of *C. elegans* primordial germ cells. *Epigenetics Chromatin* **3**, 15 (2010).
25. Sheth, U., Pitt, J., Dennis, S. & Priess, J. R. Perinuclear P granules are the principal sites of mRNA export in adult *C. elegans* germ cells. *Development* **137**, 1305–1314 (2010).
26. Hammond, T. M. et al. SAD-3, a putative helicase required for meiotic silencing by unpaired RNA, interacts with other components of the silencing machinery. *G3 (Bethesda)* **1**, 369–376 (2011).

Acknowledgements We thank members of the Kennedy laboratory for discussions. We thank T. Ishidate and C. Mello for sharing unpublished data. We thank H. Y. Mak for sharing strains. Some strains were provided by the CGC (P40 OD010440). Some strains were provided by the MITANI Laboratory. This work was supported by the National Institutes of Health, R01 GM088289 (S.K.). B.D.F. and A.S. were supported by NSF graduate research fellowships.

Reviewer information Nature thanks A. Pasquinelli and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions G.W. contributed to Figs. 1a–f, 2a–d, 3a–c, e, 4a, c and Extended Data Figs. 1a–c, 2a–c, 3a–o, 4a–c, 5a–c, 7a, c–e, 9a–i and Supplementary Fig. 1. B.D.F. contributed to Fig. 1c, d, 2a, 3b–f, 4a–c and Extended Data Figs. 1a, 6a–c, 7a–e, 8a, b, 9d, h. G.S. contributed to Extended Data Fig. 9b, c. C.P. contributed to Fig. 4a, c. A.S. contributed to Fig. 2b. S.K. supervised the project, interpreted results, contributed to Extended Data Figs. 1a and 10, and wrote the paper.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0132-0>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0132-0>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to S.K.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Strain list. N2 (WT); (NL1870) *mut-16(pk710)*; (YY009) *eri-1(mg366)*, (YY193) *eri-1(mg366)*; *nrde-2(gg91)*, (YY502) *nrde-2(gg91)*, (YY503) *nrde-2(gg90)*, (YY538) *hrde-1(tm1200)*, (YY562) *hrde-1(tm1200)*; *oma-1(zu405)*, (YY913) *nrde-2(gg518[hrde-2::3xflag::ha])*, (YY916) *znfx-1(gg544[3xflag::gfp::znfx-1])*, (YY947) *hrde-1(tm1200)*; *nrde-2(gg518)*, (YY967) *pgl-1(gg547[pgl-1::3xflag::tagrfp])*, (YY968) *znfx-1(gg544)*; *pgl-1(gg547)*, (YY996) *znfx-1(gg561)*, (TX20) *oma-1(zu405)*, (YY998) *znfx-1(gg544)*; *ego-1(gg644[ha::tagrfp::ego-1])*, (YY1020) *znfx-1(gg561)*; *oma-1(zu405)*, (SX461) *mjIS31(pie-1::gfp::h2b)*, (SS579) *pgl-1(bn101)*, (JH3225) *meg-3(tm4259)*; *meg-4(ax2026)*, (DG3226) *deps-1(bn124)*, (YY1006) *eri-1(mg366)*; *znfx-1(gg561)*, (YY1003) *eri-1(mg366)*; *znfx-1(gk458570)*, (YY1021) *znfx-1(gg561)*; *nrde-2(gg518)*, (YY1062) *znfx-1(gk458570)*, (YY1081) *deps-1(bn124)*; *mjIS31*, (YY1083) *wago-4(tm1019)*, (YY1084) *wago-4(tm2401)*, (YY1093) *wago-4(tm1019)*; *mjIS31*, (YY1094) *wago-4(tm2401)*; *mjIS31*, (YY1108) *znfx-1(gg575)*; *mjIS31*, (YY1109) *mjIS31*; *dpy-10(cn64)*, (YY1110) *eri-1(mg366)*; *wago-4(tm1019)*, (YY1111) *eri-1(mg366)*; *wago-4(tm2401)*, (YY1153) *wago-4(tm1019)*, *znfx-1(gg544)*, *mjIS31*, (YY1175) *hrde-1(gg594[ha::hrde-1])* (YY1287) *znfx-1(gg611[ha::znfx-1])*, (YY1305) *meg-3(tm4259)*; *meg-4(ax2026)*, *znfx-1(gg544)*, (YY1308) *meg-3(tm4259)*; *meg-4(ax2026)*, *pgl-1(gg547)*, (YY1325) *wago-4(gg620[3xflag::gfp::wago-4])*, (YY1326) *wago-4(gg620)*; *znfx-1(gg561)*, (YY1327) *pgl-1(gg547)*; *wago-4(gg620)*, (YY1346) *pgl-1(gg547)*; *znfx-1(gg561)*, (YY1364) *meg-3(tm4259)*; *meg-4(ax2026)*, *wago-4(gg620)*, (YY1388) *wago-4(gg627[3xflag::wago-4])*, (YY1393) *znfx-1(gg611)*; *wago-4(gg627)*, (YY1408) *znfx-1(gg575)*; *mjIS31*; *dpy-10(cn64)*, (YY1416) *znfx-1(gg544)*; *axIs1488*, (YY1419) *znfx-1(gg561)*; *wago-4(tm1019)*; *oma-1(zu405)*, (YY1442) *znfx-1(gg544)*; *hIS31397*, (YY1446) *znfx-1(gg634[ha::tagrfp::znfx-1])*, (cmp3) *mut-16::gfp::flag+loxP*, (YY1444) *znfx-1(gg634)*; *mut-16[mut-16::gfp::flag+loxP]*, (YY1452) *znfx-1(gg544)*; *hIS37(pie-1::mcherry::his58)*, (YY1453) *znfx-1(gg634)*; *wago-4(gg620)*, (YY1460) *mut-16[mut-16::gfp::flag]*; *znfx-1(gg561)*; (YY1461) *mut-16[mut-16::gfp::flag]*; *wago-4(tm1019)*; (YY1486) *znfx-1(gg631[3xflag::gfp::znfx-1 Δ helicase])*, (YY1491) *wago-4(gg620)*; *oma-1(zu405)* (YY1492) *pgl-1(gg640[pgl-1::3xflag::mcardinal])*; *mut-16[mut-16::gfp::flag+loxP]*; *znfx-1(gg634)*, (YY1494) *wago-4(tm2401)*; *pgl-1(gg547)* (YY1503) *pgl-1(gg547)*; *mut-16[mut-16::gfp::flag+loxP]*, (YY1556) *wago-4(gg627)*; *hrde-1(gg594)*.

CRISPR-Cas9. gRNAs were chosen using Ape according to following standards: first, PAM sites are in the context of GGNGG³⁹ or GNGG; second, GC content of 20-bp spacer sequence was 40% to 60%; third, high specificity according to <http://crispr.mit.edu>. All CRISPR was done using co-CRISPR strategy²⁷. Plasmids were purified with PureLink HiPure Plasmid Kits (Thermo Fisher). For deletions: two gRNAs (20 ng μl⁻¹) were co-injected into gonads with pDD162 (50 ng μl⁻¹), *unc-58* gRNA (20 ng μl⁻¹), AF-JA-76 (20 ng μl⁻¹) and 1 × taq buffer. For 3 × Flag or HA epitope tagging, single-strand oligonucleotides (4 nM ultramer from IDT, purified by isopropanol precipitation) with 50-bp homology regions were used as repair templates. gRNA (20 ng μl⁻¹) and repair template (20 ng μl⁻¹) were co-injected into gonads with pDD162 (50 ng μl⁻¹), *unc-58* gRNA (20 ng μl⁻¹), AF-JA-76 (20 ng μl⁻¹) and 1 × taq buffer. For GFP, TagRFP or mCardinal tagging, repair templates contained homologous arms of 500 bp to 1,000 bp and were cloned into pGEM-7zf(+). Sequences were confirmed by Sanger sequencing. Repair templates were amplified with PCR, gel purified and isopropanol precipitated. PCR product was heated at 95 °C for 5 min and then immediately put on ice for at least 2 min. Injection mix was prepared: pDD162 (50 ng μl⁻¹), *unc-58* gRNA (20 ng μl⁻¹), AF-JA-76 (20 ng μl⁻¹), gRNAs close to N-terminal or C-terminal of the genes (20 ng μl⁻¹), heated and cooled repair template (50 ng μl⁻¹) and 1 × standard taq buffer. Injected animals were maintained at 25 °C. Unc animals were isolated 4 days later and grown at 20 °C. Animals were screened for deletion or tagging by PCR.

RNA immunoprecipitation. Animals were flash frozen in liquid nitrogen and stored at -80 °C. Animals were resuspended in sonication buffer (20 mM Tris-HCl pH 7.5, 200 mM NaCl, 2.5 mM MgCl₂, 10% glycerol, 0.5% NP-40, 80 U ml⁻¹ RNaseOUT, 1 mM dithiothreitol (DTT) and protease inhibitor cocktail without EDTA) and sonicated (30 s on, 30 s off, 20–30% output for 2 min on a Qsonica Q880R sonicator, repeat once). Lysates were clarified by centrifuging at 18,400g for 15 min. Supernatants were precleared with protein A agarose beads and incubated with Flag-M2 agarose beads for 2–3 h at 4 °C. Beads were washed with RIP buffer (20 mM Tris-HCl pH 7.5, 200 mM NaCl, 2.5 mM MgCl₂, 10% glycerol, 0.5% NP-40) six times. Protein and associated RNAs were eluted with 100 μg ml⁻¹ 3 × Flag peptide. RNAs were treated with Turbo DNase I for 20 min at 37 °C and then extracted with TRIzol reagent followed by precipitation with isopropanol.

RT-qPCR. mRNA isolated from total RNA or from RNA immunoprecipitation experiments was converted to cDNA using the iScript cDNA synthesis kit according to vendor's instructions. The following primer sequence were used to quantify mRNA levels. *oma-1* mRNA: 5'-GCTTGAAGATATTGCATTCAACC-3'

(forward primer); 5'-AACTGTTGAAATGGAGGTGC-3' (reverse primer). *oma-1* pre-mRNA: 5'-GTGCGTTGGCTAATTCCTG-3' (forward primer); 5'-CTGAATCGCGCGAACTTG-3' (reverse primer). *gld-2* mRNA: 5'-ACGTGTAGAAAGGGCTGCAC-3' (forward primer); 5'-GTCGATGCAGATGATGATGG-3' (reverse primer). *gld-2* pre-mRNA: 5'-CCTTATTAATTCAGAGCTGCTGTC-3' (forward primer); 5'-AAGACTACACACGCAATCG-3' (reverse primer). *eft-3* pre-mRNA: 5'-CCTGCAAGTTCAACGAGCTTA-3' (forward primer); 5'-TGAAAAACAAATTTGGTACATAAAC-3' (reverse primer).

Mrt assay. Each generation, 3–6 L4 animals were picked to a single plate and grown at 25 °C; average brood sizes were calculated by counting the total number of progeny per plate.

RNAi inheritance assays. For *dpy-11* and *gfp* RNAi inheritance, embryos were collected via hypochlorite treatment and placed onto HT115 bacteria expressing dsRNA against *dpy-11* or *gfp*. F₁ embryos were collected by hypochlorite treatment from RNAi- or control-treated adults and placed onto non-RNAi plates. Worms were scored at late L4 (*dpy-11*) or early young adult (*gfp*) stages.

For *oma-1* RNAi inheritance, experiments were done at 20 °C. Embryos were collected via hypochlorite treatment and placed onto HT115 bacteria expressing dsRNA against *oma-1*. Six F₁ embryos were picked onto a single OP50 plate. From F₂ to F₆, six L4 animals were picked onto a single OP50 plate. *tm1019* is a 571-bp deletion that removes part of the PIWI domain. *tm1019* also introduces a frameshift deletion that would be expected to prevent translation of the rest of the PIWI domain. *znfx-1(gg561)* is an 8,476-bp deletion that deletes most (2,300 out of 2,400 amino acids) of ZNFX-1, including the helicase domain. Both alleles were presumably null.

Co-immunoprecipitation. Young adults were flash frozen in liquid nitrogen. Animals were ground into powder in liquid nitrogen and resuspended in 1 ml 1 × lysis buffer (20 mM HEPES pH 7.5, 100 mM NaCl, 5 mM MgCl₂, 1 mM EDTA, 10% glycerol, 0.25% Triton, 1 mM fresh-made PMSF, 1 × complete protease inhibitor from Roche without EDTA) and rotated for 45 min at 4 °C. Lysate was cleared by spinning at 2,300g for 15 min, 30 μl protein G beads were added to preclear lysate for 30 min. 3 × Flag::WAGO-4 proteins were pulled down by 30 μl agarose beads conjugated to anti-Flag antibody (A2220, Sigma-Aldrich). Input and immunoprecipitation proteins were separated by SDS-PAGE and detected by Flag M2 antibody and HA antibody (Roche, 3F10).

Small RNA sequencing. Total RNA was extracted using TRIzol. Total RNA (20 μg) was separated by 15% urea gel. Small RNA from about 18–35 nucleotides was cut from gel. Small RNAs were cloned using a 5' monophosphate independent small RNA protocol as previously described²⁸. Libraries were multiplexed with a 4-nucleotide 5' barcode and a 6-nucleotide 3' barcode and pooled for next-generation sequencing on a NextSeq 500. FastX 0.0.13 was used to separate reads that contained the 3' adaptor and filter low-quality reads for further analysis. Reads >14 nucleotides were mapped to the *C. elegans* genome (WS220) using Bowtie. Read counts were normalized to the total number of reads matching the genome. Two independent libraries were prepared and the two replicates were combined for Fig. 1f.

Microscopy and analysis. To image larval and adult stages, animals were immobilized in M9 with 0.1% sodium azide, and mounted on glass slides. To image embryos, gravid adults were dissected on a coverslip containing 10 μl of 1 × egg buffer, and then mounted on freshly made 3% agarose pads. Animals were imaged immediately with a Nikon Eclipse Ti microscope equipped with a W1 Yokogawa Spinning disk with 50 μm pinhole disk and an Andor Zyla 4.2 Plus sCMOS monochrome camera. A 60 ×/1.4 Plan Apo Oil objective was used unless otherwise stated. *pie-1::gfp::h2b* imaging was done using a widefield Zeiss Axio Observer.Z1 microscope equipped with a Plan-Apochromat 63 ×/1.40 Oil DIC M27 objective and an ORCA-Flash 4.0 CMOS Camera.

Colocalization. The degree of colocalization between different fluorescently labelled proteins across development was calculated using the Coloc2 plugin from ImageJ. Animals were imaged as described above with the exception of using a 100 ×/1.45 Plan Apo Oil objective. Around 3–5 granules were selected from at least 3 different animals across each stage of development specified. Region of interest (ROI) masks were generated using the 3D ROI Manager plugin in ImageJ to eliminate black regions surrounding granules²⁹. Coloc2 was used to generate a Pearson's R value for degree of colocalization between two channels in the region defined by the ROI mask.

FRAP. FRAP experiments were conducted using a Zeiss LSM 780 point scanning confocal equipped with a Quasar PMT x2 + GAASP 32 Channel Spectral Detector using a 63 ×/1.4 Plan Apo Oil objective. Adult animals (for pachytene germ cells) or embryos (for P₂ blastomere) were suspended in a mixture of 0.5% sodium azide and 50% 0.1 μm polystyrene beads (Polysciences) to inhibit movement. The mixture was added to a coverslip and placed on a fresh 3% agarose pad. Slides were sealed with nail polish. The bleaching plugin within the Zeiss Black software was used to specify the ROI to be bleached. One ROI was used for all data points.

Single z-slice images were acquired at 1 s intervals for 15 s, followed by bleaching, then continued at 1 s intervals for 85 s. Images were aligned using neighbouring granules in ImageJ to account for subtle shifts in movement. An ROI was generated around the bleached region and continuously measured across all time points using the plot profile function within ImageJ. Data were normalized to an unbleached control granule to account for background bleaching throughout the 100-s period. Normalized data points were averaged across all seven granules and plotted using Prism. The heat map of a representative granule was generated using the thermal LUT within ImageJ.

Quantification of distances between foci centres and surfaces. We imaged pachytene germ cell nuclei in three animals. Approximately ten granules were selected from each animal. Confocal z stacks were opened with the 3D objects counter plugin from ImageJ to generate x, y and z coordinates for the centre of each object³⁰. To account for chromatic shift between channels, 0.1 µm Tetraspek beads were imaged and granule distances were corrected accordingly. Distances between foci surfaces was calculated with 3D ROI manager in ImageJ²⁹. Thresholding function within 3D ROI manager was used to eliminate background signal.

Immunofluorescence. Approximately 30 animals were sliced open in 8 µl of 1 × egg buffer (25 mM HEPES, pH 7.3, 118 mM NaCl₂, 48 mM KCl, 2 mM CaCl₂, 2 mM MgCl₂) to isolate gonads and embryos. A coverslip was added and slides were placed on a metal block (chilled on dry ice) for 10 min. Coverslips were popped off and slides were submerged in methanol at −20 °C for 10 min, followed by acetone at −20 °C for 5 min. Samples were allowed to dry at room temperature for 3 min. Then, 500 µl of PBS with Tween 20 (1 × PBST) was added to each sample and incubated for 5 min at room temperature followed by 500 µl of 1 × PBST and 1% bovine serum albumin (BSA) for 30 min at room temperature. Antibody solution (50 µl of 1 × PBST, 1% BSA, 1:20 dilution of anti PGL-1 antibody (K76 from DSHB), and 1:250 dilution of anti-HA antibody (abcam ab9110)) was added to each sample. Samples were covered with parafilm and incubated overnight at room temperature inside a humid chamber. Samples were washed three times in 1 × PBST at room temperature for 10 min. Secondary antibodies (Alexa Fluor 555 goat anti-rabbit, Life Technologies A21429; and Alexa Fluor 488 goat anti-mouse, Life Technologies A10667) were diluted 1:50 in 1 × PBST. Secondary solution (50 µl) was added to each sample, covered with parafilm, and incubated for 90 min in the dark at room temperature. Samples were washed three times in 1 × PBST at room temperature for 10 min. Vectashield antifade (15 µl) and DAPI was added to each sample. Slides were sealed with nail polish.

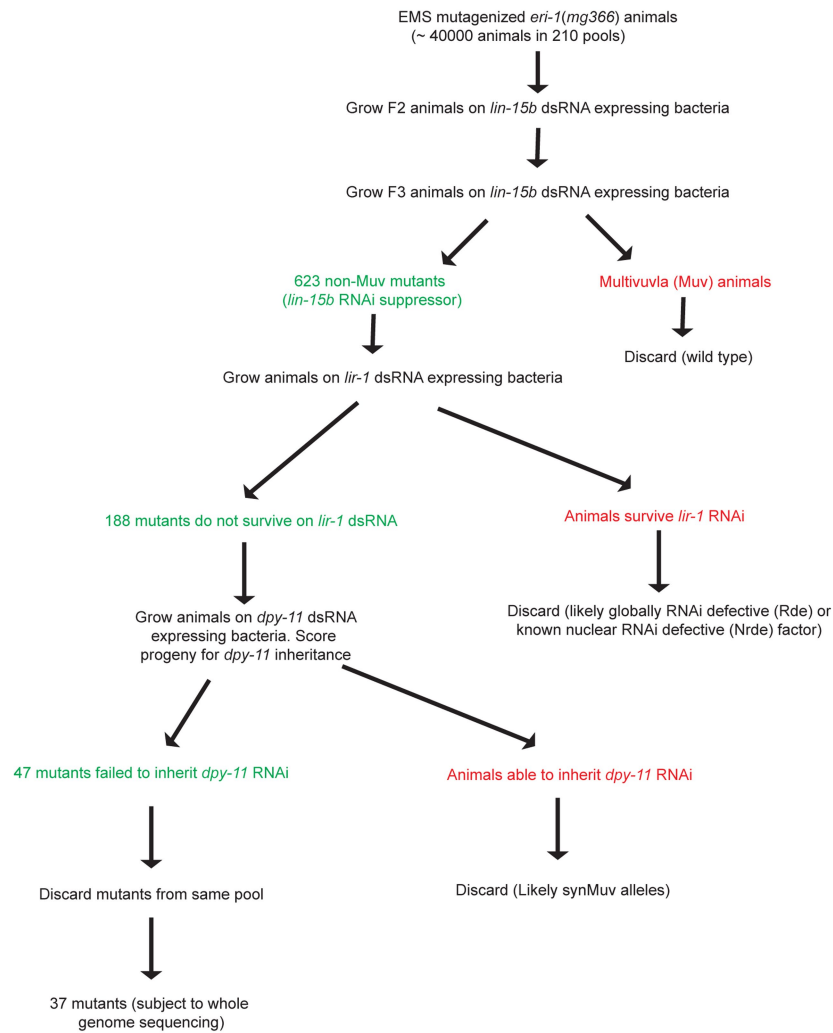
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. Custom python scripts used to generate small RNA plots in Fig. 1f are available upon request.

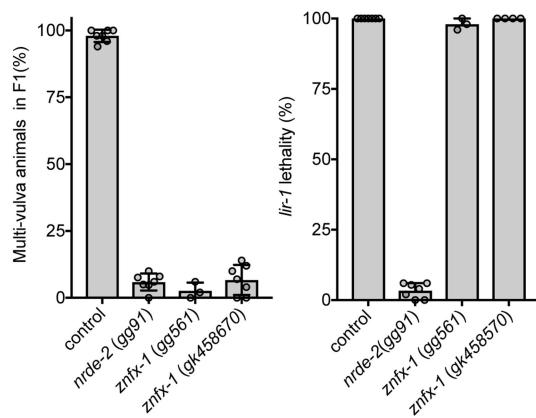
Data availability. Small RNA sequencing data that support the findings of this study have been deposited in the Expression Omnibus (GEO) database with the accession code GSE112109. Source data for Fig. 2b, d is located in Supplementary File 1. The remaining data that support the findings of this study are available from the corresponding author upon reasonable request.

27. Arribere, J. A. et al. Efficient marker-free recovery of custom genetic modifications with CRISPR/Cas9 in *Caenorhabditis elegans*. *Genetics* **198**, 837–846 (2014).
28. Gent, J. I. et al. A *Caenorhabditis elegans* RNA-directed RNA polymerase in sperm development and endogenous RNA interference. *Genetics* **183**, 1297–1314 (2009).
29. Ollion, J., Cochenne, J., Loll, F., Escudé, C. & Boudier, T. TANGO: a generic tool for high-throughput 3D image analysis for studying nuclear organization. *Bioinformatics* **29**, 1840–1841 (2013).
30. Bolte, S. & Cordelières, F. P. A guided tour into subcellular colocalization analysis in light microscopy. *J. Microsc.* **224**, 213–232 (2006).
31. Blumenthal, T. et al. A global analysis of *Caenorhabditis elegans* operons. *Nature* **417**, 851–854 (2002).
32. Clark, S. G., Lu, X. & Horvitz, H. R. The *Caenorhabditis elegans* locus lin-15, a negative regulator of a tyrosine kinase signaling pathway, encodes two different proteins. *Genetics* **137**, 987–997 (1994).
33. Huang, L. S., Tzou, P. & Sternberg, P. W. The lin-15 locus encodes two negative regulators of *Caenorhabditis elegans* vulval development. *Mol. Biol. Cell* **5**, 395–411 (1994).
34. Guang, S. et al. An Argonaute transports siRNAs from the cytoplasm to the nucleus. *Science* **321**, 537–541 (2008).
35. Burton, N. O., Burkhart, K. B. & Kennedy, S. Nuclear RNAi maintains heritable gene silencing in *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **108**, 19683–19688 (2011).
36. Lin, R. A gain-of-function mutation in *oma-1*, a *C. elegans* gene required for oocyte maturation, results in delayed degradation of maternal proteins and embryonic lethality. *Dev. Biol.* **258**, 226–239 (2003).
37. Kawasaki, I. et al. PGL-1, a predicted RNA-binding component of germ granules, is essential for fertility in *C. elegans*. *Cell* **94**, 635–645 (1998).
38. Spike, C. A., Bader, J., Reinke, V. & Strome, S. DEPS-1 promotes P-granule assembly and RNA interference in *C. elegans* germ cells. *Development* **135**, 983–993 (2008).
39. Farboud, B. & Meyer, B. J. Dramatic enhancement of genome editing by CRISPR/Cas9 through improved guide RNA design. *Genetics* **199**, 959–971 (2015).

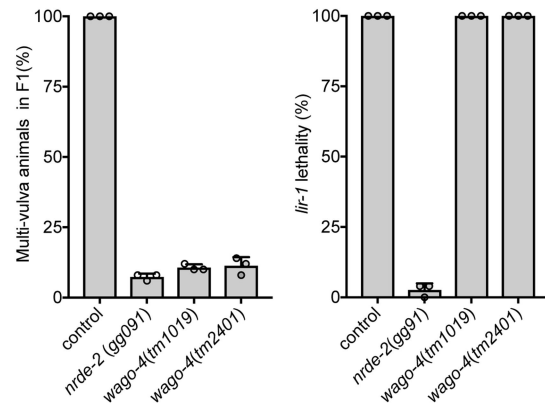
a



b



c

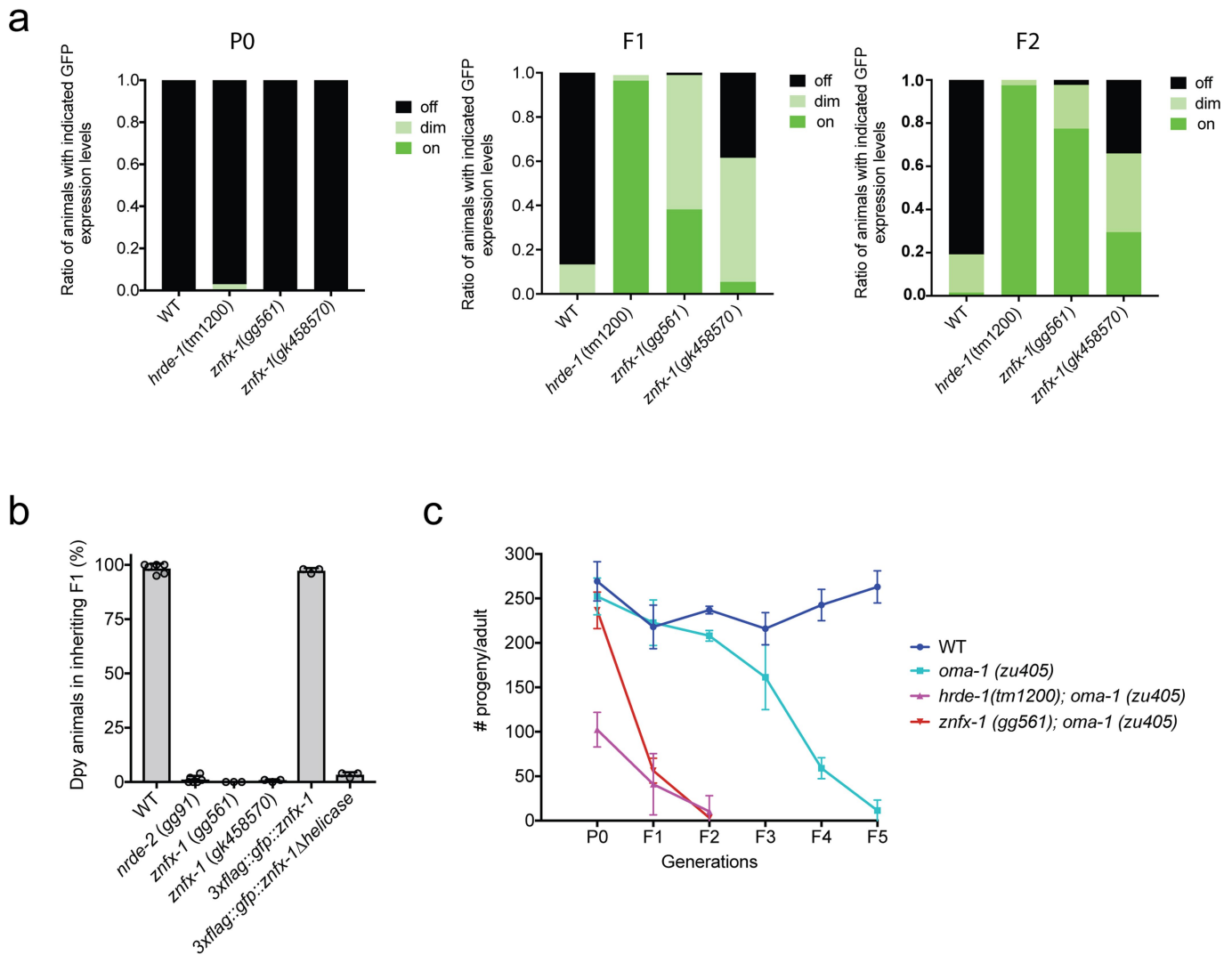


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Genetic screen to identify novel RNAi

inheritance mutants. a, A genetic screen was conducted to identify components of the *C. elegans* RNAi inheritance machinery. The screen contained several filters (see below) to remove known RNAi inheritance factor. Factors defective for RNAi inheritance are also defective for nuclear RNAi⁶. Therefore, our screen began with selections for mutant alleles that disrupt nuclear RNAi. Two selections were developed for nuclear RNAi mutants. First, the *lin-15b* and *lin-15a* genes are transcribed as a polycistronic message that is spliced within the nucleus into *lin-15b* and *lin-15a* mRNAs³¹. Animals containing mutations in both *lin-15b* and *lin-15a* exhibit a multivulva (Muv) phenotype^{32,33}. RNAi targeting *lin-15b* (in *eri-1*(–) animals) silences *lin-15b* and *lin-15a* co-transcriptionally, thus inducing a Muv phenotype³⁴. The previously identified nuclear RNAi factors are required for *lin-15b* RNAi-induced co-transcriptional silencing of *lin-15a* and, therefore, for *lin-15b* RNAi-induced Muv. A second assay for nuclear RNAi is *lir-1* RNAi. *lir-1* RNAi is lethal because *lir-1* is in an operon with *lin-26*, and co-transcriptional silencing of *lin-26* by *lir-1* RNAi causes lethality³⁴. Nuclear RNAi defective (NRDE) animals do not die in response to *lir-1* RNAi because they fail to silence *lin-26*³⁴. Previous genetic screens have used suppression of *lir-1* RNAi to find factors required for nuclear RNAi. These screens have reached saturations: we have identified several alleles in all the *nrde* genes using this approach. Unpublished work from the laboratory shows, however, that hypomorphic alleles of the *nrde* genes will often block *lin-15b* RNAi-induced Muv and yet still die in response to *lir-1* RNAi. We interpret these data to mean that survival from *lir-1* RNAi is a much stronger selection for nuclear RNAi mutants than a failure to form Muv in response to *lin-15b* RNAi. That is,

factors that contribute to nuclear RNAi, but are not 100% required for nuclear RNAi, would not be identified by *lir-1* RNAi suppression screens. Therefore, our screen looked for suppressors of *lin-15b* RNAi, which did not suppress *lir-1* RNAi, because this screen might identify genes missed in our previous genetic screens. Step 1, identify factors required for nuclear RNAi. *eri-1*(*mg366*) animals were mutagenized with EMS. F₂ progeny were exposed to bacteria expressing *lin-15b* dsRNA. Non-Muv animals were kept as candidate novel nuclear RNAi factors. Step 2, discard known nuclear RNAi factors. We probably know all non-essential genes that can mutate to suppress *lir-1* RNAi. Therefore, we discarded mutants that suppressed *lir-1* RNAi as these alleles are probably known nuclear RNAi factors. Mutants that did not suppress *lir-1* may contain mutations in factors important, but not essential, for nuclear RNAi. Step 3, identify mutations that suppress RNAi inheritance. The last filter in our screen was to identify mutant alleles that disrupted RNAi inheritance. We subjected remaining mutant animals to *dpy-11* RNAi, which causes animals to become Dumpy (Dpy). Progeny of animals exposed to *dpy-11* dsRNA inherit *dpy-11* silencing and are Dpy³⁵. RNAi inheritance mutants become Dpy in response to *dpy-11* RNAi; however, the progeny of these animals fail to inherit *dpy-11* silencing, and, therefore, are not Dpy. Thus, any of our mutant animals that became Dpy in response to *dpy-11* RNAi, but whose progeny were not Dpy, were kept for further analysis. Finally, only one mutant was kept from each pool (pools were maintained as independent populations throughout the screen). **b, c,** Independent alleles of *znfx-1* and *wago-4* are (as expected) defective for *lin-15b* RNAi and not defective for *lir-1* RNAi. Data are mean \pm s.d. of more than three biologically independent samples.

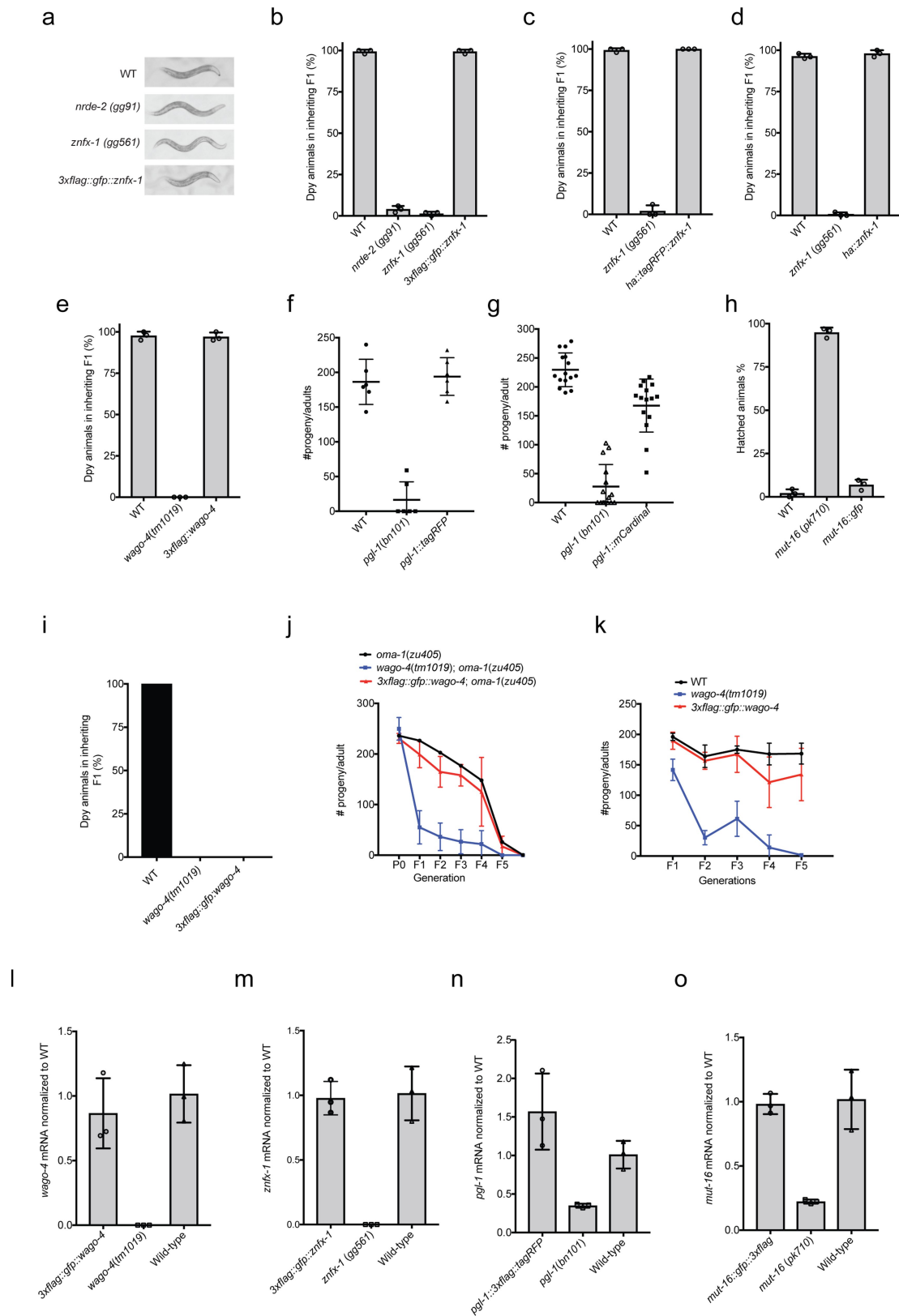


Extended Data Fig. 2 | ZNFX-1 is required for RNAi inheritance.

a, Animals expressing a *pie-1::gfp::h2b* transgene were exposed to *gfp* dsRNA⁴. The percentage of the P₀, F₁ and F₂ progeny of the indicated genotypes expressing GFP was quantified. Data represent scoring of at least 80 animals in each generation and for each genotype. Note, the *gfp* reporter transgene used in this study is a multi-copy version of the single copy version used in Fig. 1b. Note that some RNAi inheritance can be seen in *znfx-1* mutant animals using this reporter transgene. Thus, in some cases, some RNAi inheritance can occur in the absence of ZNFX-1.

b, Animals of the indicated genotypes were exposed to *dpy-11* dsRNA. The F₁ progeny of these animals were grown in the absence of *dpy-11*

dsRNA, and were scored for Dpy phenotypes. Data are mean \pm s.d. of more than three biologically independent samples. Consistent with the idea that ZNFX-1 (and NRDE-2) is required specifically for inheritance, *znfx-1* mutant animals exposed directly to *dpy-11* dsRNA are Dpy (data not shown). **c**, *zu405ts* is a temperature-sensitive (ts) lethal (embryonic arrest at 20°C) allele of *oma-1*³⁶. *oma-1* RNAi suppresses *oma-1(zu405ts)* lethality, and this effect is heritable^{5,6}. Animals of the indicated genotypes were exposed to *oma-1* dsRNA and the fertility of the progeny of these animals was scored over generations. Data show that *znfx-1* mutant animals are defective for *oma-1* RNAi inheritance. Data are mean \pm s.d. of three biologically independent samples.

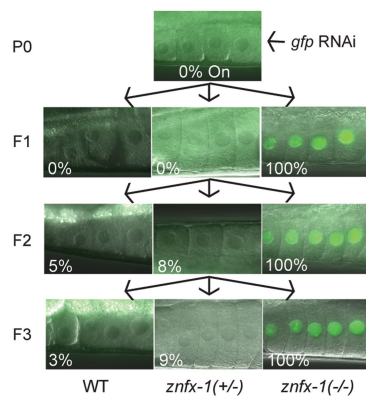


Extended Data Fig. 3 | See next page for caption.

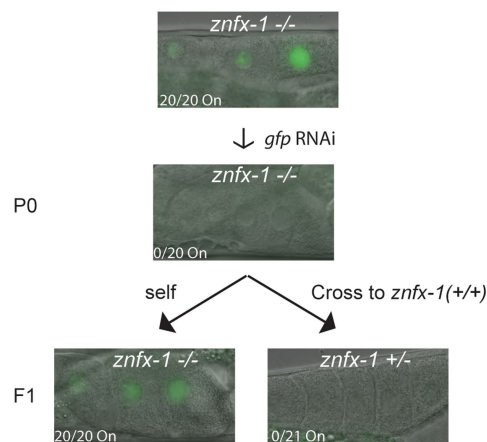
Extended Data Fig. 3 | CRISPR–Cas9-epitope tagged genes used in this study, with one exception, produce functional proteins and are expressed at or near wild-type levels. a–e, The addition of epitope tags by CRISPR–Cas9-mediated gene conversion of *znfx-1* or *wago-4* did not affect function of tagged proteins in these RNAi inheritance. *dpy-11* RNAi inheritance assays in which the progeny of animals exposed to *dpy-11* dsRNA are visually scored for the inheritance of Dpy phenotypes. The indicated epitope-tagged proteins are functional in this RNAi inheritance assay. $n = 3$ biologically independent samples; data are mean \pm s.d. **f, g,** *pgl-1* mutant animals show a temperature-sensitive (25 °C) sterile phenotype. The addition of epitope tags by CRISPR–Cas9-mediated gene conversion to the *pgl-1* locus did not affect PGL-1 function as these animals were fertile. L4 animals were singled from 20 °C to 25 °C and brood sizes were scored. *pgl-1::tagrfp* ($n = 6$ animals) and *pgl-1::mcardinal; tagrfp::znfx-1*; *mut-16::gfp* ($n = 15$ animals). **h,** *mut-16(–)* animals are defective for *pos-1* RNAi. Embryos of the indicated genotype were grown on *pos-1* dsRNA-expressing bacteria. Six L4 animals were picked to *pos-1* dsRNA-expressing bacteria and laid eggs overnight. Unhatched embryos and hatched animals were scored. The addition of *gfp* to the *mut-16* locus did not affect MUT-16 function. Data are mean \pm s.d. of three biologically independent samples. **i–k,** In some cases, Flag::GFP::WAGO-4-expressing animals are defective in RNAi inheritance, indicating that Flag::GFP::WAGO-4 is

not fully functional. **i,** Animals of the indicated genotypes were exposed to *dpy-11* dsRNA and F₁ progeny were grown in the absence of *dpy-11* dsRNA. The percentage of Dpy animals is shown. At least 150 animals of each genotype were scored. Thus, 3 \times Flag::GFP::WAGO-4 is not functional for *dpy-11* inheritance. $n = 3$ biologically independent samples. **j,** 3 \times Flag::GFP::WAGO-4 is functional during *oma-1* RNAi inheritance. See Extended Data Fig. 2c for details of the *oma-1* RNAi inheritance assay. $n = 3$ biologically independent samples; data are mean \pm s.d. In Fig. 2d, both *wago-4* and *znfx-1* are shown to exhibit an Mrt phenotype at 25 °C. Here, 3 \times flag::gfp::wago-4 animals are not Mrt, indicating that 3 \times Flag::GFP::WAGO-4 is capable of promoting germline immortality. $n = 3$ biologically independent samples; data are mean \pm s.d. **i–o,** CRISPR tags did not seem to affect gene expression. To address the possibility that epitope tagging of the genes used in this study changed gene expression levels, we isolated total RNA from animals of the indicated genotypes and used qRT-PCR to quantify indicated mRNA levels. Primers target exon–intron junctions. Early stop or deletion alleles for each of these loci were used as controls. *wago-4(tm1019)* and *znfx-1(gg561)* are deletions and primers were located within deleted regions. *pgl-1(bn101)* and *mut-16(pk710)* are nonsense alleles. A decrease in the mRNA levels of these mutants is probably due to nonsense-mediated decay. $n = 3$ biologically independent samples; data are mean \pm s.d.

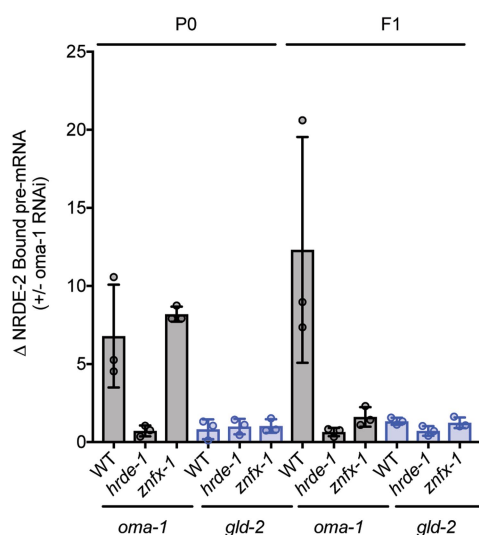
a



b

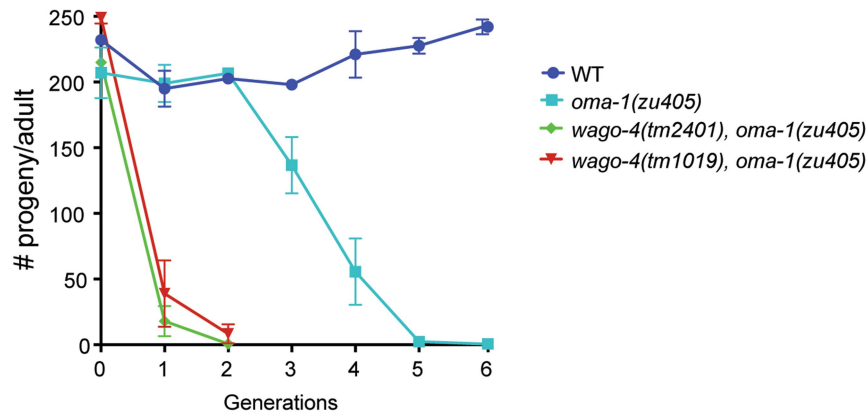


c

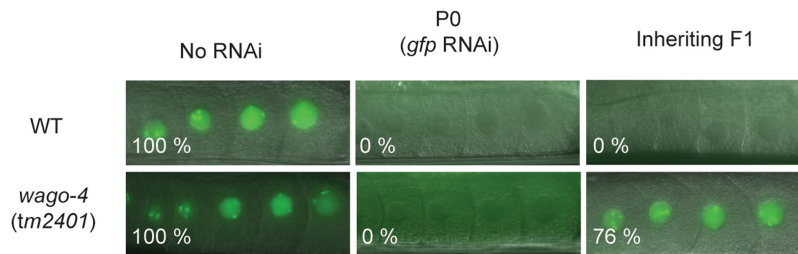


Extended Data Fig. 4 | ZNFX-1 acts specifically during the inheriting phase of RNAi. **a**, *znfx-1* is required in inheriting generations for RNAi inheritance to occur. In brief, we initiated gene silencing in *znfx-1/+* heterozygous animals and scored the *+/+* and *-/-* progeny for their ability to inherit gene silencing. Progeny containing at least one wild-type copy of *znfx-1* were capable of inheriting gene silencing, whereas *-/-* progeny were not. More specifically, *znfx-1(gg575) +/-* animals that express the *pie-1::gfp::h2b* transgene¹⁷ exposed to *gfp* dsRNA, and progeny from F₁ to F₃ generations were scored. Micrographs of GFP fluorescence in oocytes are shown. To identify cross progeny, the following strategy was used via CRISPR. *pie-1::gfp::h2b* was marked by *dpy-10(cn64)* (*dpy-10* is approximately 0.77 cM from *pie-1::gfp::h2b*). *dpy-10(cn64)/+* animals are Dpy Rol and *dpy-10(cn64)* homozygous animals are Dpy. *znfx-1* genotypes were inferred based upon wild-type, Dpy and Rol phenotypes; *n* > 30 animals. **b**, *znfx-1* is sufficient in inheriting generations for RNAi inheritance to occur. We initiated gene silencing in *znfx-1(-/-)* animals, introduced a wild-type copy of *znfx-1* to progeny (via mating), and scored *znfx-1/+* cross-progeny for inheritance. The data show that *znfx-1(+/-)* progeny, from parents that lack a wild-type copy of *znfx-1*, were able to inherit silencing. *znfx-1(gg575)* was marked by *dpy-10(cn64)* (*dpy-10* is approximately 1.09 cM from *znfx-1*). *dpy-10(cn64)/+* animals are Dpy Rol and *dpy-10(cn64)* homozygous animals are Dpy. *znfx-1* genotypes was inferred based upon wild-type, Dpy and Rol phenotypes. *n* > 20 animals. **c**, Additional biochemical evidence that ZNFX-1 acts in inheriting generations to promote inheritance. The nuclear RNAi factor NRDE-2 binds to pre-mRNA of genes undergoing heritable silencing⁶. When *znfx-1(-)* animals were exposed directly to *oma-1* dsRNA, NRDE-2 bound to the *oma-1* pre-mRNA at wild-type levels. However, in progeny of *znfx-1(-)* mutant animals NRDE-2 failed to bind *oma-1* pre-mRNA. Animals expressing NRDE-2::3×Flag were treated with *oma-1* RNAi. Extracts were generated from these animals as well as the progeny of these animals (which were not treated directly with *oma-1* RNAi). NRDE-2::3×Flag was immunoprecipitated with an anti-Flag antibody and NRDE-2 co-precipitating *oma-1* pre-mRNA was quantified by qRT-PCR with exon-intron primer sets designed to detect unspliced RNAs (pre-mRNAs) of the *oma-1* gene as well as a control germline expressed pre-mRNA *gld-2*. *hrde-1* allele *tm1200* and *znfx-1* allele *gg561* were used. Data are mean ± s.d. of the ratio of signals ± *oma-1* RNAi; *n* = 3 biological replicates.

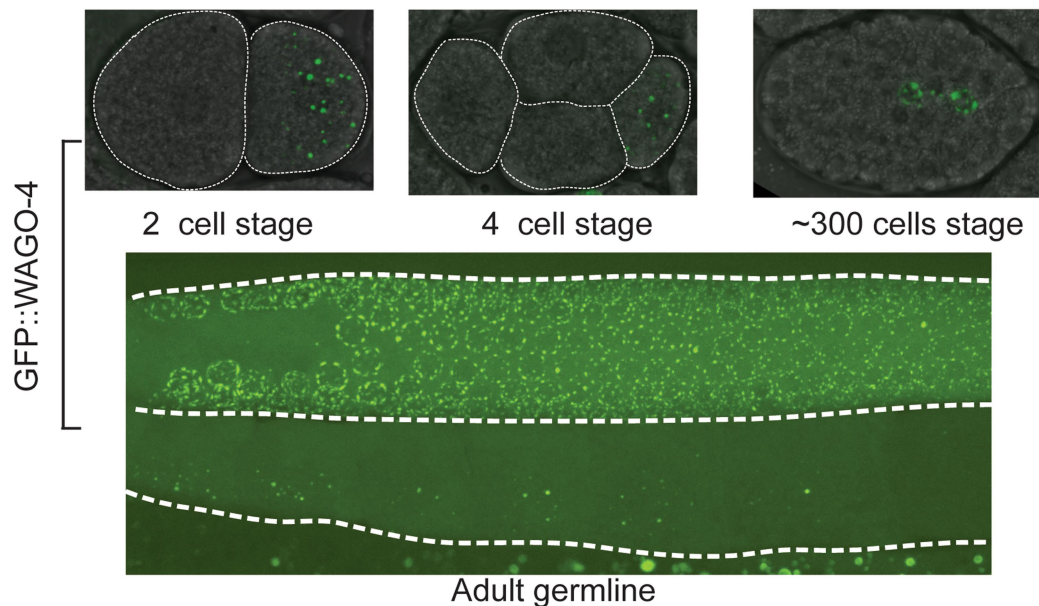
a



b

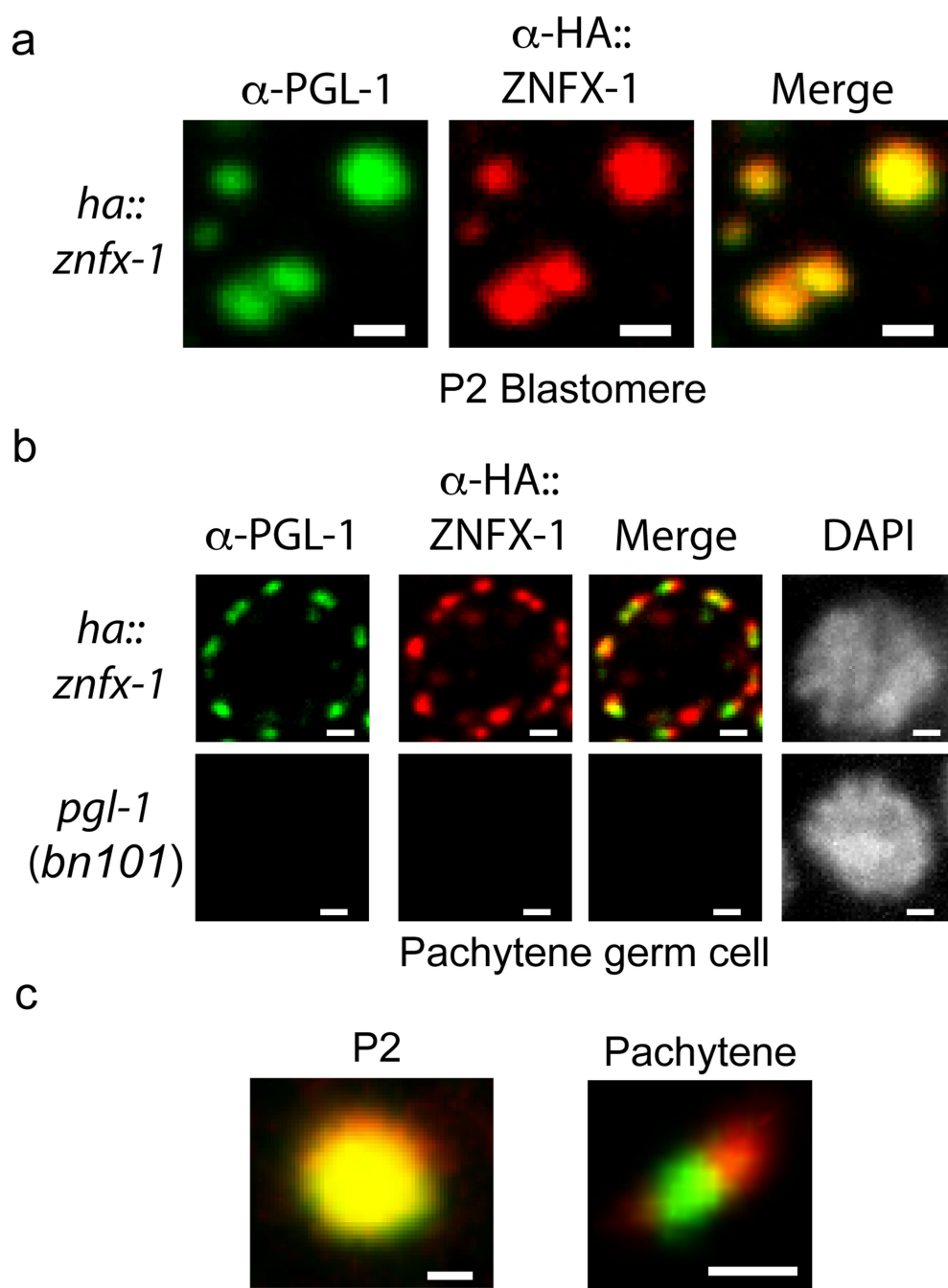


c



Extended Data Fig. 5 | WAGO-4 is an Argonaute that localizes to the peri-nucleus and is required for RNAi inheritance. **a**, *oma-1(zu405)* is a temperature-sensitive lethal (embryonic arrest at 20°C) allele of *oma-1*. *oma-1* RNAi suppresses *oma-1(zu405)* lethality and this effect is heritable⁵. Animals of the indicated genotypes were exposed to *oma-1* dsRNA, and F₁ to F₅ progeny were grown in the absence of *oma-1* dsRNA. Number of viable progeny of P₀ (directly exposed to *oma-1* RNAi) and inheriting generations (F₁ to F₅, grown in the absence of *oma-1* RNAi) were scored (20°C). Data are mean ± s.d. of three biologically independent samples. **b**, Animals of the indicated genotypes and expressing a *pie-1::gfp::h2b*

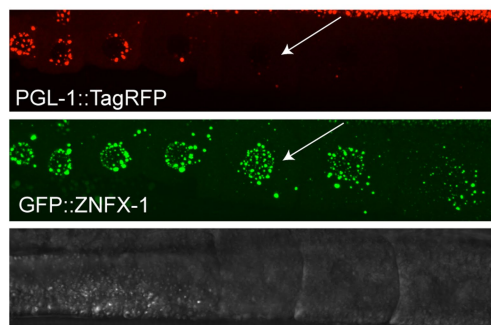
transgene were exposed to *gfp* dsRNA¹⁷. Micrographs of animals +/– *gfp* RNAi as well as the F₁ progeny of these animals are shown. The percentage of animals expressing GFP is indicated, and represent the scoring of at least 90 animals in each generation and for each genotype. **c**, We used CRISPR–Cas9 to append a *gfp* tag upstream of the predicted *wago-4* atg start codon. Top, fluorescent micrographs of *gfp::wago-4* in 2-cell, 4-cell and ~300-cell embryos. Bottom, fluorescent micrograph of the germline of an adult *gfp::wago-4* animal. Images are representative of more than three animals at each lifestage.



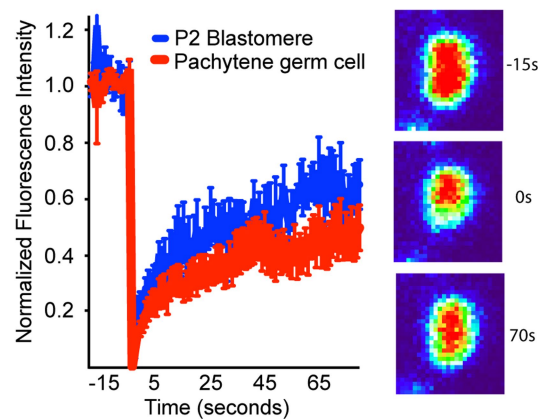
Extended Data Fig. 6 | Visualization of Z granule formation with antibodies targeting PGL-1 (P granule) and HA::ZNFX-1 (Z granule). To control for possible artefacts caused by fluorescent epitopes, we conducted immunofluorescence on HA::ZNFX-1-expressing animals using anti-PGL-1 (K76 Developmental Studies Hybridoma Bank) and anti-HA (Abcam ab9110) antibodies. **a**, Anti-PGL-1 and anti-HA signals colocalized in the P₂ blastomeres of 4-cell embryos. **b**, Anti-PGL-1

and anti-HA signals were adjacent, yet distinct in, in pachytene germ cells. No PGL-1 or HA::ZNFX-1 signal was detected in *pgl-1(bn101)* animals, which do not express PGL-1 or HA::ZNFX-1, establishing that immunofluorescent signals were specific. **c**, Magnification of foci from **a** and **b**. Images in **a–c** are representative of three independent animals at each life stage. Scale bars, 1 μ m (**a**), 1 μ m (**b**) and 0.5 μ m (**c**).

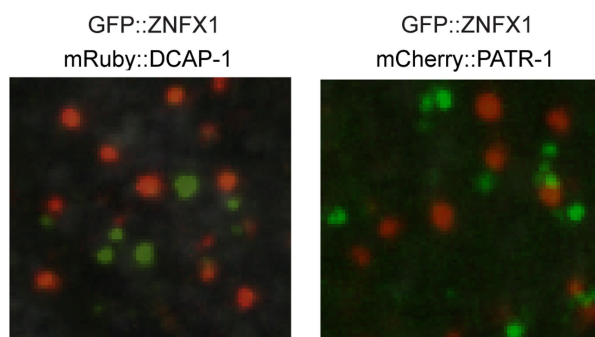
a



b

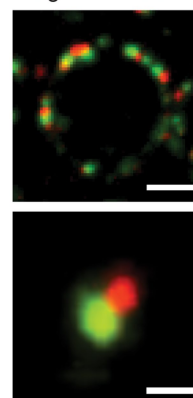


c

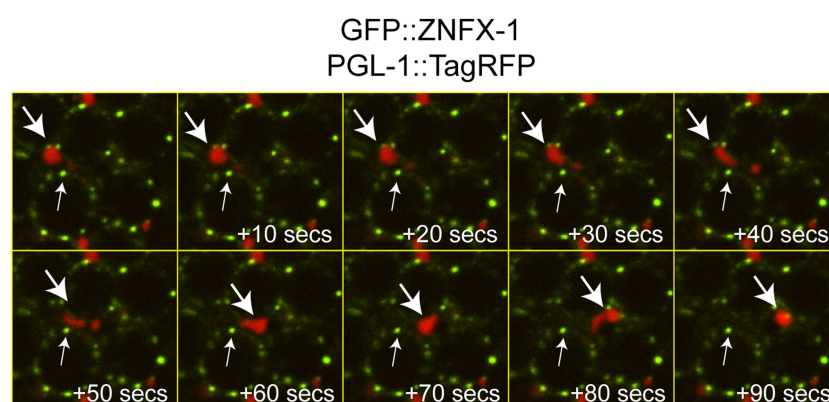


d

GFP::ZNFX-1
TagRFP::EGO-1



e

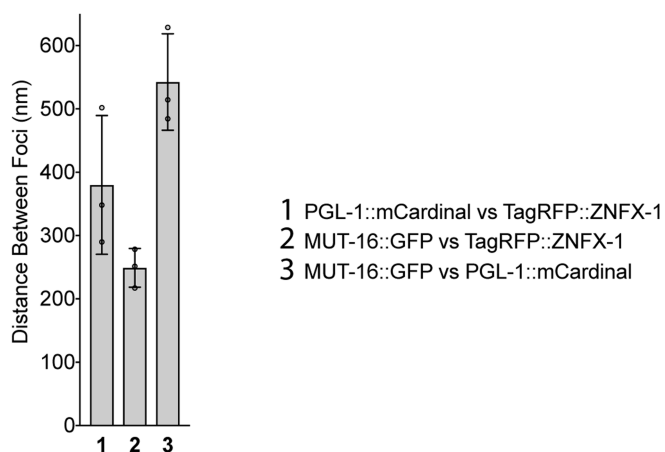


Extended Data Fig. 7 | See next page for caption.

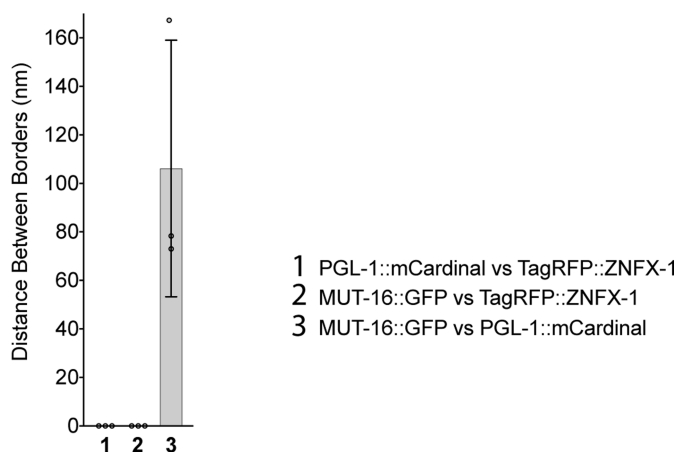
Extended Data Fig. 7 | Z granules independently form liquid-like condensates that do not colocalize with P bodies but localize adjacent to EGO-1 foci, and can be physically and temporally dissociated from P granules. **a**, During oocyte maturation, ZNFX-1 foci detach from the nuclei, assume spherical shapes and move away from the nucleus; this behaviour is consistent with Z foci being liquid-like condensates. In addition, the data show that Z foci can exist at developmental stages during which P granules are no longer visible, indicating that Z foci can be temporally separated from P granules. Image is of maturing oocytes of animals expressing the indicated fluorescent proteins. Long arrows indicate oocytes that contain Z granules, but not P granules. Image representative of more than three animals. **b**, Z foci exhibit properties reminiscent of liquid droplets. Left, GFP::ZNFX-1-expressing animals were subjected to FRAP (see Methods) and fluorescence was monitored in bleached area over indicated time. Data are normalized to a non-bleached control granule from the same sample. Data are mean \pm s.e.m. of $n = 7$ individual granules from 7 animals. Right, heat maps showing recovery of ZNFX-1 fluorescence in a representative bleached Z granule. **c**, Z foci do not colocalize with other known liquid droplets. GFP::ZNFX-1 does not colocalize with markers of processing bodies. PATR-1 and DCAP-1 localize to processing bodies¹⁵. Fluorescent micrographs of somatic

blastomeres of embryos expressing the indicated fluorescent proteins. ZNFX-1 does not colocalize with markers of processing bodies in these cells. Images are representative of more than three independent animals. **d**, Z foci in adult pachytene germ cells do not colocalize with EGO-1. ZNFX-1 foci form adjacent to EGO-1 foci. Fluorescent micrographs of a single pachytene germ cell nucleus from animals expressing GFP::ZNFX-1 and TagRFP::EGO-1. A 3D render of a representative foci is shown below. Images are representative of three independent animals. Scale bars, 0.5 μ m. **e**, Z foci can be physically separated from P granules. Gonads were isolated from animals expressing GFP::ZNFX-1 and PGL-1::TagRFP and subjected to shearing force as described⁸. Time-lapse imaging at 10-s intervals is shown. A PGL-1-labelled P granule detaching from the nucleus and flowing throughout the cytoplasm is shown (large arrow). ZNFX-1-labelled Z granules remain immobile (small arrow). Physical shearing was induced as previously described⁸. In brief, GFP::ZNFX-1 and PGL-1::TagRFP adults were dissected to extrude gonads. Isolated gonads were squeezed between two coverslips to generate shearing force. Coverslips were then mounted on a slide and imaged immediately with a spinning disc confocal. Z stacks were acquired every 10 s. Images are representative of four independent animals.

a

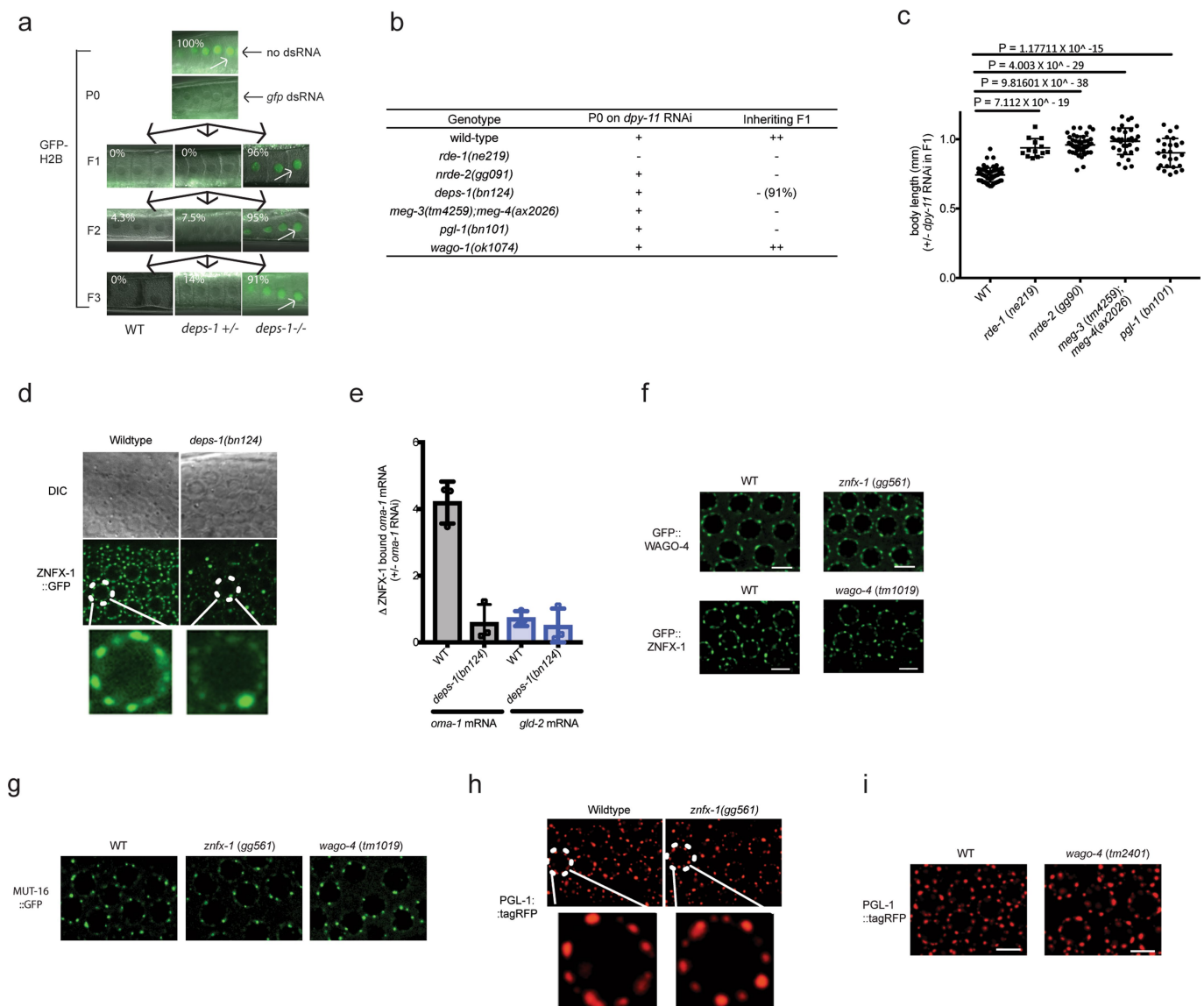


b



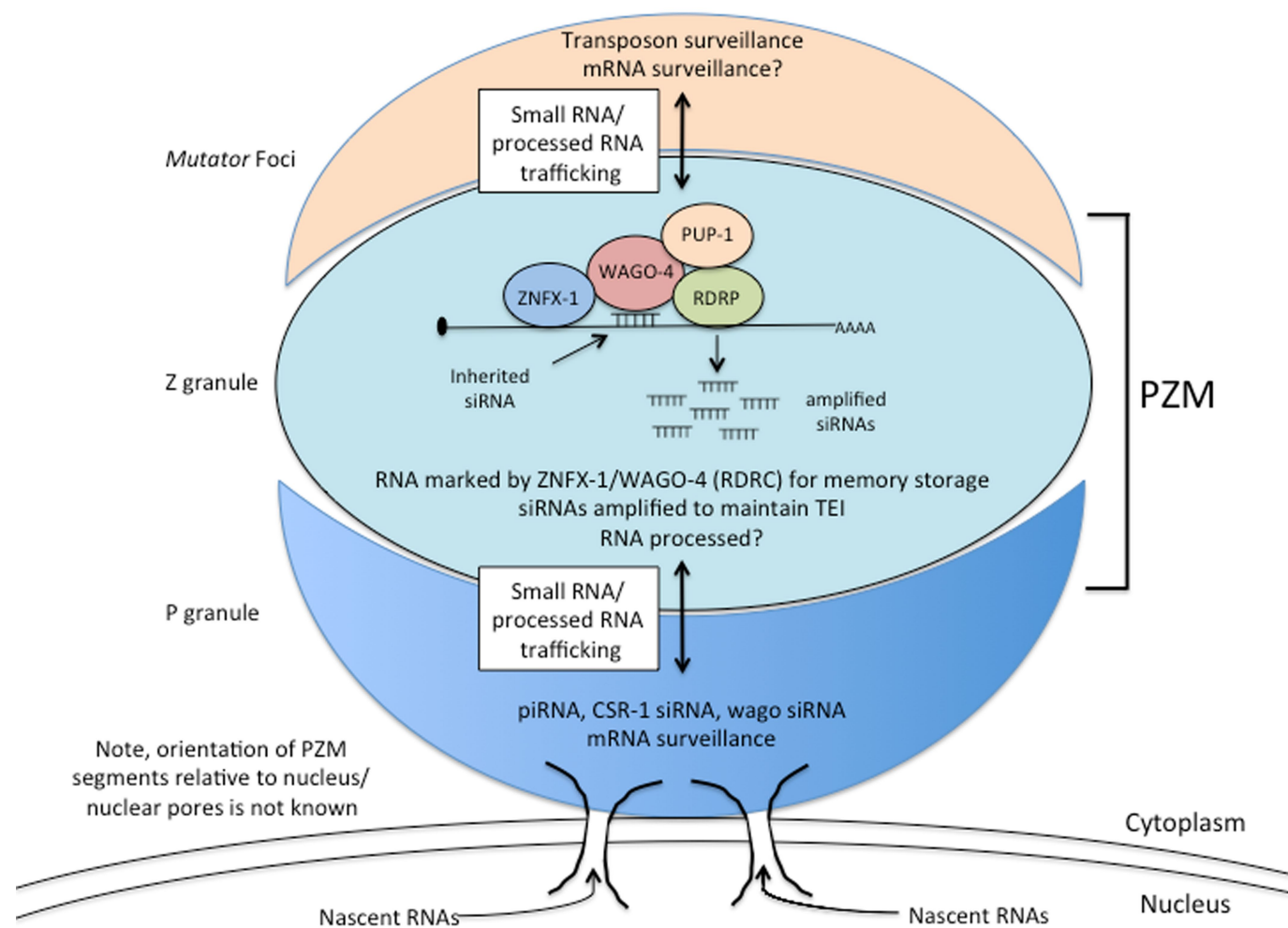
Extended Data Fig. 8 | Quantification of centres and surfaces of fluorescence for Z granules, P granules and *Mutator* foci. a, b, Distances between centres (a) and surfaces (b) of the spaces occupied by PGL::mCardinal, TagRFP::ZNFX-1 and MUT-16::GFP were calculated

as described in the Methods. Data are mean \pm s.d. of 10 granule measurements across 3 independent animals. Distances have been corrected for chromatic shift.



Extended Data Fig. 9 | P granule assembly factors contribute to RNAi inheritance, normal Z granule morphology and the ability of ZNFX-1 to bind mRNAs. a, DEPS-1 is required for P granule formation in adult germ cells^{11,37,38}. *deps-1(bn124)/+* animals expressing the *pie-1::gfp::h2b* transgene¹⁷ were exposed to *gfp* dsRNA. Progeny were grown in the absence of *gfp* dsRNA for three generations. Fluorescent micrographs show GFP expression in oocytes. The percentage of animals expressing *pie-1::gfp::h2b* is shown. These data show that DEPS-1 activity is required in inheriting generations to allow for *gfp* RNAi inheritance. $n = 32$ animals for P₀, $n > 100$ for F₁ to F₃. **b**, MEG-3/4, DEPS-1 and PGL-1 also contribute to P granule formation^{11,37,38}. *dpy-11* RNAi causes animals exposed to *dpy-11* dsRNA to become Dumpy (Dpy). Progeny of animals exposed to *dpy-11* dsRNA inherit *dpy-11* silencing and are Dpy³⁵. RNAi inheritance mutants become Dpy in response to *dpy-11* RNAi; however, progeny fail to inherit *dpy-11* silencing, and, therefore, are not Dpy. Animals of indicated genotypes were exposed to *dpy-11* dsRNA. F₁ progeny were grown in the absence of *dpy-11* dsRNA. (–) indicates non-Dpy; (+) indicates mild Dpy phenotype; (++) indicates strong Dpy. *pgl-1*, *deps-1*, and *meg-3/4* are defective for *dpy-11* RNA inheritance. $n = 3$ biologically independent samples for each condition. **c**, Animals of indicated genotypes were exposed to *dpy-11* dsRNA and F₁ progeny were grown in the absence of *dpy-11* dsRNA. Body lengths of F₁ animals were measured by Image J.

Data are expressed as body length from progeny of *dpy-11* RNAi-treated animals divided by the average body length from control animals. The mean of $n > 12$ animals with P values calculated by Student's two-tailed t -test is shown. **d**, In *deps-1(bn124)* animals, most Z granules are smaller than normal while one Z granule/nucleus becomes enlarged. Images are from pachytene region of germline. Images are representative of more than three animals. **e**, In *deps-1(bn124)* animals, ZNFX-1 does not bind *oma-1* dsRNA. ZNFX-1 was immunoprecipitated in RNAi generation with anti-Flag antibodies and co-precipitating RNA was subjected to qRT-PCR to quantify *oma-1* mRNA co-precipitating with ZNFX-1 in wild-type or *deps-1(bn124)* animals. *gld-2* is a germline-expressed control mRNA. Data are mean \pm s.d. of three biologically independent samples. **f–i**, Loss of ZNFX-1 or WAGO-4 does not seem to affect the formation of Z granules marked by GFP::WAGO-4 or GFP::ZNFX-1 (**f**), Mutator foci marked by MUT-16::GFP (**g**), or P granules marked by PGL-1::RFP (**h**, **i**). Note, in late embryonic germline development, PGL-1::TagRFP foci may not be efficiently concentrated into Z2/Z3 in *wago-4* mutant (data not shown). Images are representative of more than three animals. All images in **d–i** were taken using a 60 \times objective, and scaled to the same size as the other images within a panel. Scale bars, 5 μ m (**f**, **i**).



Extended Data Fig. 10 | Working model for role of PZM assemblages in RNAi inheritance. P granules make contacts with nuclear pores^{23,25}.

The relationship between the Z and M segments of the PZM granule and nuclear pores are not yet known.

A small amount of mini-charged dark matter could cool the baryons in the early Universe

Julian B. Muñoz^{1*} & Abraham Loeb²

The dynamics of our Universe is strongly influenced by pervasive—albeit elusive—dark matter, with a total mass about five times the mass of all the baryons^{1,2}. Despite this, its origin and composition remain a mystery. All evidence for dark matter relies on its gravitational pull on baryons, and thus such evidence does not require any non-gravitational coupling between baryons and dark matter. Nonetheless, some small coupling would explain the comparable cosmic abundances of dark matter and baryons³, as well as solving structure-formation puzzles in the pure cold-dark-matter models⁴. A vast array of observations has been unable to find conclusive evidence for any non-gravitational interactions of baryons with dark matter^{5–9}. Recent observations by the EDGES collaboration, however, suggest that during the cosmic dawn, roughly 200 million years after the Big Bang, the baryonic temperature was half of its expected value¹⁰. This observation is difficult to reconcile with the standard cosmological model but could be explained if baryons are cooled down by interactions with dark matter, as expected if their interaction rate grows steeply at low velocities¹¹. Here we report that if a small fraction—less than one per cent—of the dark matter has a mini-charge, a million times smaller than the charge on the electron, and a mass in the range of 1–100 times the electron mass, then the data¹⁰ from the EDGES experiment can be explained while remaining consistent with all other observations. We also show that the entirety of the dark matter cannot have a mini-charge.

A new arena for the search for interactions between dark matter (DM) and baryons can be found at the cosmic dawn. During this era, the first stars were formed¹², and their ultraviolet emission coupled the spin temperature of neutral hydrogen to the much lower kinetic temperature^{13,14}, causing cosmic-microwave-background (CMB) photons with a local wavelength of 21 cm to be resonantly absorbed by the intervening neutral hydrogen. Eventually, the baryonic gas was heated by X-ray sources, and the hydrogen spin temperature increased above that of the CMB, triggering 21-cm emission¹⁵. The absorption era, however, provides one of the lowest-velocity environments in our Universe: DM interactions with the visible sector mediated by a massless field, such as the photon, are expected to be at their most prominent at that time. We will use the idea of the low-velocity environment to explore the possibility that the DM interacts with baryons through a small electric charge.

Endowing the DM with a ‘mini-charge’ has unique phenomenological consequences, as charged particles respond to background magnetic fields, which are rather common in astrophysical environments. It was argued that supernova shocks would eject all mini-charged particles from the Galactic Disk¹⁶, and that the Galactic magnetic field, known¹⁷ to extend beyond Galactic heights of 3 kpc, would prevent them from re-entering the disk. Given that the DM density within 1.5 kpc of the disk is in agreement with predictions¹⁸, we conclude that not all DM can be evacuated from the disk, and thus, mini-charged particles with charges larger than $\epsilon/m_\chi \gtrsim 5 \times 10^{-16} \text{ MeV}^{-1}$ are precluded from being the entirety of the cosmological DM. Throughout this paper, we define the DM mini-charge ϵ in units of the electron charge e , so $\epsilon \equiv e_\chi/e$, where e_χ is the DM charge. A comparable constraint can be obtained from studying galaxy clusters,

which we detail in the Methods section. We note that, even though the precise numerical value of these limits on ϵ/m_χ can be altered by different assumptions, the charge required for DM to cool the baryons would be orders of magnitude larger.

These constraints would not apply if only a fraction of the DM is charged, as most of the DM would behave as expected. Given that the local DM measurements are accurate to within tens of per cent, we will focus on the possibility that the mini-charged particles constitute a small fraction $f_{\text{dm}} \leq 0.1$ of the DM, while the rest of it is neutral. This can be naturally achieved if DM forms ‘dark atoms’, with a small charged-DM fraction remaining free after its recombination¹⁹, although we will posit no assumptions about the origin of the mini-charged particles. The momentum-transfer cross-section between a mini-charged particle and a target t (electron or proton) is²⁰

$$\bar{\sigma}_t = \frac{2\pi c^2 \hbar^2 \alpha^2 \epsilon^2 \xi}{\mu_{\chi,t}^2 v^4} \quad (1)$$

where $\mu_{\chi,t}$ is the reduced mass of the target and DM, v is the relative velocity between the two particles, α is the fine-structure constant, c the speed of light, \hbar the reduced Planck constant and ξ the Debye logarithm⁷, which we compute in the Methods section. The velocity behaviour of this cross-section is that of Rutherford scattering, growing as the DM–baryon fluid becomes slower and, by extension, colder.

The relative velocity between the DM and baryons is not determined by their thermal motion alone. The gravitational infall of baryons is impeded until hydrogen recombination occurs, whereas the DM streams freely, causing a velocity difference between them²¹. Interactions between DM and baryons cause a drag on this velocity²², which eventually leads to mechanical equilibrium of the two fluids, dissipating the velocity into thermal energy. Additionally, interactions between DM and baryons will tend to bring the two fluids into thermal equilibrium, equating their temperatures. Given that the DM is very cold, this can severely lower the baryon temperature. To reduce the baryonic temperature substantially (as is required to explain the EDGES data¹⁰) with DM–baryon interactions, equipartition demands the existence of at least as many mini-charged particles as baryons, which translates into a DM mass $m_\chi \leq \bar{\mu}_p f_{\text{dm}} (\Omega_c/\Omega_b)$, where $\bar{\mu}_p$ is the mean molecular weight of baryons, and Ω_c and Ω_b are the cosmic abundances of DM and baryons, respectively. Thus, for each value of f_{dm} we will study only the mass range $m_\chi \leq 6.2 \text{ GeV} \times f_{\text{dm}}$.

We solve for the thermodynamical evolution of the DM and baryonic fluids simultaneously, accounting for their relative velocity, as well as the small free-electron fraction left after recombination^{22,23} (see Methods section for details). This yields the baryonic temperature $T_b(z, v_{\chi,b}^{(i)})$ as a function of the initial DM–baryon relative velocity, $v_{\chi,b}^{(i)}$. To remove dependencies on the astrophysics of the coupling between the spin and kinetic temperatures, we define the average baryonic temperature as

$$\langle T_b(z) \rangle = \int dv_{\chi,b}^{(i)} \mathcal{P}(v_{\chi,b}^{(i)}) T_b(z, v_{\chi,b}^{(i)}) \quad (2)$$

¹Department of Physics, Harvard University, Cambridge, MA, USA. ²Astronomy Department, Harvard University, Cambridge, MA, USA. *e-mail: julianmunoz@fas.harvard.edu

where the probability distribution function of the initial velocity, $\mathcal{P}(v)$, is given by a Maxwell–Boltzmann distribution with root mean square (r.m.s.) velocity $v_{\text{rms}} = 29 \text{ km s}^{-1}$ at decoupling²¹.

Figure 1 shows, for different values of f_{dm} , the lines in the ϵ – m_χ plane that would produce enough baryonic cooling to explain the EDGES measurement¹⁰. Given f_{dm} , the required mini-charge scales as $\epsilon \propto m_\chi$. There is, however, no simple analytic solution for the slope of this line as a function of f_{dm} , since for small energy transfers we expect the baryon heating to be $\dot{Q}_b \propto f_{\text{dm}} \epsilon^2/m_\chi^2$, whereas for large energy transfers (and assuming $f_{\text{dm}} < 0.2$), $\dot{Q}_b \propto f_{\text{dm}}^{5/2} \epsilon^2/m_\chi^2$. We have empirically found that

$$\epsilon(m_\chi, f_{\text{dm}}) \approx 6 \times 10^{-7} \left(\frac{m_\chi}{\text{MeV}} \right) \left(\frac{f_{\text{dm}}}{10^{-2}} \right)^{-3/4} \quad (3)$$

is sufficient to reduce the baryonic temperature by a factor of 2, although we emphasize that the 21-cm results in Fig. 1 have been calculated numerically for each value of f_{dm} .

Next, we summarize the relevant constraints on mini-charged DM. Minicharged-particle production during the supernova 1987A would have altered its neutrino luminosity^{24,25}, thus constraining the range $10^{-7} < \epsilon < 10^{-9}$, which we label SN1987A in Fig. 1. A search for mini-charged particles at SLAC National Accelerator Laboratory²⁶ placed constraints on mini-charges larger than $\epsilon \approx 10^{-4}$ for $m_\chi < 100 \text{ MeV}$. We show this constraint in Fig. 1 labelled as SLAC mQ. Measurements of the matter power spectrum, from the CMB and the Lyman- α forest, can only constrain mini-charged particles if they compose a major part of the DM⁷. Otherwise, even particles with mini-charges $\epsilon \gtrsim 10^{-6} (m_\chi/\text{MeV})^{0.3}$, which would be in thermal contact with baryons at the CMB epoch, are allowed to compose up to 1% of the DM²⁷. This constraint, nonetheless, closes the apparent gap for $m_\chi \gtrsim 200 \text{ MeV}$ in Fig. 1, as, above this threshold, the 21-cm data would require more than 1% of the DM to have a charge. Thus, we will focus on the $f_{\text{dm}} \lesssim 10^{-2}$ range for the rest of this paper.

The cosmology of mini-charged particles can place additional constraints on their charge. Particles with mini-charges larger than $\epsilon \gtrsim 10^{-8} (m_\chi/\text{MeV})^{1/2}$, which encompasses the region of interest, would reach equilibrium with the visible sector in the early Universe. This places limits on mini-charged particles lighter than electrons, since they

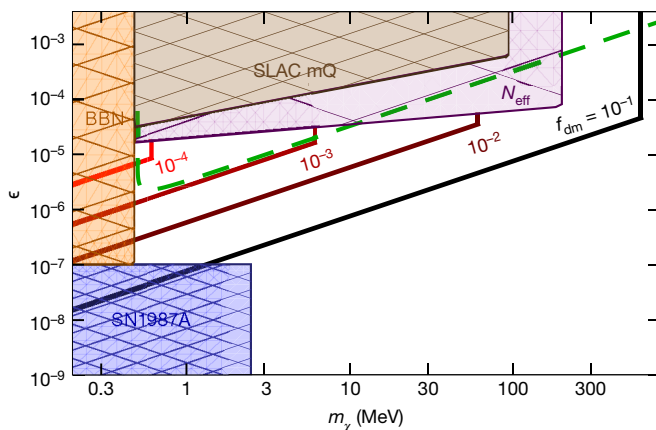


Fig. 1 | Regions of the mini-charged-particle parameter space explored by 21-cm observations, and current constraints. Each thick solid line (from black to red) represents the mini-charge required to reduce the baryonic temperature by a factor of two, as reported by EDGES¹⁰, if a fraction f_{dm} of the DM is mini-charged. We require $m_\chi \leq 6.2 \text{ GeV} \times f_{\text{dm}}$ to produce enough cooling (with lines ending abruptly at that mass), and $f_{\text{dm}} < 10^{-1}$, as larger values are ruled out by other observations. The coloured hatched regions are excluded by different datasets, and the green long-dashed line represents the mini-charge required to obtain the appropriate DM abundance.

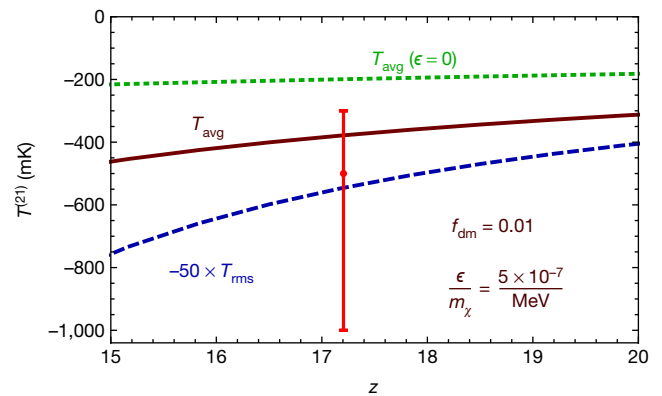


Fig. 2 | Brightness temperature of 21-cm emission as a function of redshift. We assume full Lyman- α coupling and no X-ray heating, both in the case with (solid line) and without (dotted line) DM–baryon interactions. Negative temperatures indicate absorption. The red data point represents the data from EDGES¹⁰, of $T^{(21)} = -500^{+200}_{-500} \text{ mK}$, at 3σ (99.73% confidence levels shown by error bar). We also show (dashed blue line) the r.m.s. of the 21-cm temperature due to velocity fluctuations, multiplied by a factor of -50 .

would appear as additional light degrees of freedom (parametrized as a change in the effective number N_{eff} of neutrino species) during Big Bang nucleosynthesis (BBN)²⁸. We label the constrained region in Fig. 1 as BBN. Moreover, if a dark photon is the origin of the mini-charge, it can also alter BBN and the CMB through the same mechanism²⁹. We estimate this constraint in the Methods section, and label it as N_{eff} in Fig. 1. In the standard freeze-out scenario, the DM production is halted when the baryonic temperature drops below its mass, and its annihilation rate determines the relic abundance left in the dark sector. We compute the mini-charge required to produce the appropriate DM abundance and plot it as the dashed line in Fig. 1. It is clear that—barring a small region for m_χ of a few MeV, and $\epsilon \gtrsim 10^{-6}$ —most of the parameter space that we are considering is below this thermal-relic line, thus requiring new interactions to allow the DM to annihilate efficiently. We leave this challenge for future model building of the needed dark sector.

Let us now study how a change in baryonic temperature translates into an observable 21-cm brightness temperature. The 21-cm temperature is inversely proportional to the gas temperature during the cosmic dawn, and shown in Fig. 2 for a specific choice of ϵ/m_χ and f_{dm} . The EDGES data¹⁰ is in tension with the maximum absorption possible in the standard model, whereas this tension is resolved when introducing mini-charged DM particles. Thus, we conclude that a subpercent fraction of the DM with an electric charge $e_\chi \approx 10^{-6} e$ and mass about 1–60 MeV can cool the baryons considerably, while being consistent with all current constraints. This scenario predicts new inhomogeneities in the baryon temperature, since the DM–baryon relative velocity, with fluctuations over Mpc scales²¹, modulates the overall cooling/heating, thus forming a source of additional 21-cm fluctuations²². We can estimate the size of these fluctuations by finding the root-mean-squared 21-cm brightness temperature, as a function of the DM–baryon velocity. Figure 2 shows that the same interactions that cause baryonic cooling also lead to additional 21-cm fluctuations at the level of a few per cent. These are comparable to the Mpc-scale adiabatic fluctuations at $z \approx 17$, and are potentially detectable with upcoming 21-cm interferometers, such as HERA³⁰. Their detection would confirm that DM and baryons were in thermal contact during the cosmic dawn, and would thus constitute an indication of DM physics beyond the standard model.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0151-x>.

Received: 21 December 2017; Accepted: 20 February 2018;
Published online 30 May 2018.

- Ade, P. A. R. et al. Planck 2015 results. XIII. Cosmological parameters. *Astron. Astrophys.* **594**, A13 (2015).
- Rubin, V. C., Ford, W. K. Jr & Thonnard, N. Rotational properties of 21 SC galaxies with a large range of luminosities and radii, from NGC 4605/ $R = 4$ kpc/ to UGC 2885/ $R = 122$ kpc/. *Astrophys. J.* **238**, 471–487 (1980).
- Kaplan, D. B. A Single explanation for both the baryon and dark matter densities. *Phys. Rev. Lett.* **68**, 741–743 (1992).
- Weinberg, D. H., Bullock, J. S., Governato, F., Kuzio de Naray, R. & Peter, A. H. G. Cold dark matter: controversies on small scales. *Proc. Natl Acad. Sci. USA* **112**, 12249–12255 (2015).
- Akerib, D. S. et al. Limits on spin-dependent WIMP–nucleon cross section obtained from the complete LUX exposure. *Phys. Rev. Lett.* **118**, 251302 (2017).
- Ackermann, M. et al. The Fermi Galactic Center GeV excess and implications for dark matter. *Astrophys. J.* **840**, 43 (2017).
- Dvorkin, C., Blum, K. & Kamionkowski, M. Constraining dark matter–baryon scattering with linear cosmology. *Phys. Rev. D* **89**, 023519 (2014).
- Fox, P. J., Harnik, R., Kopp, J. & Tsai, Y. LEP shines light on dark matter. *Phys. Rev. D* **84**, 014028 (2011).
- Muñoz, J. B. & Loeb, A. Constraints on dark matter–baryon scattering from the temperature evolution of the intergalactic medium. *J. Cosmol. Astropart. Phys.* **1711**, 043 (2017).
- Bowman, J., Rogers, A. E. E., Monsalve, R. A., Mozdzen, T. J. & Mahesh, N. An absorption profile centred at 78 megahertz in the sky-averaged spectrum. *Nature* **555**, 67–70 (2018).
- Barkana, R. Possible interaction between baryons and dark-matter particles revealed by the first stars. *Nature* **555**, 71–74 (2018).
- Loeb, A. & Furlanetto, S. R. *The First Galaxies in the Universe* (Princeton Univ. Press, Princeton, 2013).
- Wouthuysen, S. A. On the excitation mechanism of the 21-cm (radio-frequency) interstellar hydrogen emission line. *Astron. J.* **57**, 31–32 (1952).
- Field, G. B. The spin temperature of intergalactic neutral hydrogen. *Astrophys. J.* **129**, 536–550 (1959).
- Furlanetto, S., Oh, S. P. & Briggs, F. Cosmology at low frequencies: the 21 cm transition and the high-redshift universe. *Phys. Rep.* **433**, 181–301 (2006).
- Chuzhoy, L. & Kolb, E. W. Reopening the window on charged dark matter. *J. Cosmol. Astropart. Phys.* **0907**, 014 (2009).
- Jansson, R. & Farrar, G. R. A new model of the galactic magnetic field. *Astrophys. J.* **757**, 14 (2012).
- Bovy, J. & Tremaine, S. On the local dark matter density. *Astrophys. J.* **756**, 89 (2012).
- Cline, J. M., Liu, Z. & Xue, W. Millicharged atomic dark matter. *Phys. Rev. D* **85**, 101302 (2012).
- McDermott, S. D., Yu, H.-B. & Zurek, K. M. Turning off the lights: how dark is dark matter? *Phys. Rev. D* **83**, 063509 (2011).
- Tselikhovich, D. & Hirata, C. Relative velocity of dark matter and baryonic fluids and the formation of the first structures. *Phys. Rev. D* **82**, 083520 (2010).
- Muñoz, J. B., Kovetz, E. D. & Ali-Haïmoud, Y. Heating of baryons due to scattering with dark matter during the dark ages. *Phys. Rev. D* **92**, 083528 (2015).
- Ali-Haïmoud, Y. & Hirata, C. M. HyRec: A fast and highly accurate primordial hydrogen and helium recombination code. *Phys. Rev. D* **83**, 043513 (2011).
- Davidson, S., Hannestad, S. & Raffelt, G. Updated bounds on millicharged particles. *J. High Energy Phys.* **05**, 003 (2000).
- Essig, R. et al. *Working Group Report: New Light Weakly Coupled Particles*. Available at <https://arxiv.org/abs/1311.0029> (2013).
- Prinz, A. A. et al. Search for millicharged particles at SLAC. *Phys. Rev. Lett.* **81**, 1175–1178 (1998).
- Dolgov, A. D., Dubovsky, S. L., Rubtsov, G. I. & Tkachev, I. I. Constraints on millicharged particles from Planck data. *Phys. Rev. D* **88**, 117701 (2013).
- Jaekel, J. & Ringwald, A. The low-energy frontier of particle physics. *Annu. Rev. Nucl. Part. Sci.* **60**, 405–437 (2010).
- Vogel, H. & Redondo, J. Dark radiation constraints on mini-charged particles in the universe with a hidden photon. *J. Cosmol. Astropart. Phys.* **1402**, 029 (2014).
- DeBoer, D. R. et al. Hydrogen Epoch of Reionization Array (HERA). *Publ. Astron. Soc. Pacif.* **129**, 045001 (2017).

Acknowledgements We thank P. Agrawal, Y. Ali-Haïmoud, C. Dvorkin, M. Kamionkowski, D. Pinner, C. Stubbs and S. Westerdale for discussions. This research is supported in part by the Black Hole Initiative, which is funded by a JTF grant.

Author contributions J.B.M. performed the calculations and wrote the code, with assistance from A.L. Both authors wrote the manuscript.

Competing interests The authors declare no competing financial interests.

Additional information

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to J.B.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Cluster constraints. Additional constraints on mini-charged particles can be achieved by requiring that the DM must not be trapped in regions of coherent magnetic field within galaxy clusters³¹, which have a typical correlation lengths r_{corr} of the order of 10 kpc and field strengths $B \approx 5 \mu\text{G}$. This means that charged particles with charges larger than $\epsilon/m_\chi \gtrsim 3 \times 10^{-17} \text{ MeV}^{-1}$ would not be distributed as cold DM but instead clump wherever magnetic fields are coherent. Additional constraints can be derived, through plasma effects in cluster collisions, such as the bullet cluster³², as well as by requiring the mini-charged particles not to diffuse within clusters³³, although simulations would be required to isolate these effects from nonlinear gravity.

Debye logarithm. Here, and throughout the Methods section, we work in natural units, where $\hbar = c = 1$, and also set the Boltzmann constant to unity. In the main text we defined the Debye logarithm, ξ , which regulates the forward divergence of the momentum-transfer integral²⁰. This factor is roughly constant during the era of interest, so we will set it to

$$\xi = \log \left(\frac{9T_b^3}{4\pi\epsilon^2 \alpha^3 x_e n_H} \right) \approx 68 - 2 \log \left(\frac{\epsilon}{10^{-6}} \right) \quad (4)$$

where we adopt a fiducial baryonic temperature $T_b = 10 \text{ K}$, and evaluate the free-electron fraction x_e and the number density n_H of hydrogen nuclei at redshift $z = 20$.

Drag and heating terms. DM–baryon interactions cause a drag $D(v_{\chi,b}) \equiv dv_{\chi,b}/dt$, on their relative velocity²², which we recast as

$$D(v_{\chi,b}) = \sum_{t=e,p} \frac{\sigma_t m_\chi n_\chi + \rho_b m_t n_H}{m_\chi + m_t} \frac{F(r_t)}{\rho_b v_{\chi,b}} \quad (5)$$

where ρ_b is the baryon energy density. The number density of mini-charged DM is given by $n_\chi = f_{\text{dm}} \rho_d / m_\chi$, where ρ_d is the (total) DM energy density at redshift z , and m_t is the target mass. Here we have defined the function

$$F(r_t) = \text{erf} \left(\frac{r_t}{\sqrt{2}} \right) - \sqrt{\frac{2}{\pi}} r_t e^{-r_t^2/2} \quad (6)$$

where $r_t \equiv v_{\chi,b} / u_{\text{th},t}$ and the thermal sound speed of the DM–target fluid is given by

$$u_{\text{th},t} = \sqrt{\frac{T_b}{m_t} + \frac{T_\chi}{m_\chi}} \quad (7)$$

where T_χ is the mini-charged-DM temperature. By comparing this sound speed with the relative velocity, we see that in the standard case of $T_\chi = 0$, immediately after recombination (and before the X-ray heating of the baryons), the baryonic sound speed falls below the DM–baryon relative velocity, making the DM–proton fluid (albeit not the DM–electron fluid) ‘supersonic’.

In addition to damping the DM–baryon relative velocity, these interactions give rise to a baryonic heating²²,

$$\begin{aligned} \dot{Q}_b = n_\chi \frac{x_e}{1 + f_{\text{He}}} \sum_{t=e,p} \frac{m_\chi m_t}{(m_\chi + m_t)^2} \frac{\sigma_t}{u_{\text{th},t}} \\ \times \left[\sqrt{\frac{2}{\pi}} \frac{e^{-r_t^2/2}}{u_{\text{th},t}^2} (T_\chi - T_b) + m_\chi \frac{F(r_t)}{r_t} \right] \end{aligned} \quad (8)$$

where $f_{\text{He}} \equiv n_{\text{He}}/n_H \approx 0.08$. Here, we have included DM interactions with both protons and electrons, as the latter can dominate if the DM fluid is not cold. The DM heating can be found by symmetry, through the transformations $n_\chi \rightarrow n_H$, $m_\chi \leftrightarrow m_t$ and $T_\chi \leftrightarrow T_b$. This heating can be positive or negative depending on r_t , as for $r_t \rightarrow 0$ (corresponding to $v_{\chi,b} \ll u_{\text{th},t}$) only the temperature-dependent term survives, corresponding to the usual thermalization³⁴; whereas for $r_t \gg 1$ (which implies $v_{\chi,b} \gg u_{\text{th},t}$), the heating term proportional to $F(r_t)$ dominates, converting the mechanical energy of the relative velocity into heat for both fluids.

Dark-matter sound speed. For illustration purposes, we note that transferring half of the baryonic thermal energy to the DM at $z = z_{\text{central}}$ would induce a DM sound-speed of

$$u_\chi^2 = \frac{T_\chi}{m_\chi} \approx \frac{T_b}{\mu_b f_{\text{dm}} \Omega_c} \approx \frac{(0.1 \text{ km s}^{-1})^2}{f_{\text{dm}}} \quad (9)$$

at $z = z_{\text{central}} \approx 17$, where the EDGES data lie¹⁰. Interestingly, for $f_{\text{dm}} \lesssim 0.2$, this gives the mini-charged component of the DM a sound speed larger than that of protons, thus setting the value of $u_{\text{th},p}$ in equation (7). Moreover, we estimate that

for $f_{\text{dm}} \lesssim 10^{-2}$ the DM–electron interactions will dominate over the DM–proton ones, owing to this velocity. This would, however, suppress the matter power spectrum only on extremely small scales and for a minority of the DM.

Thermal evolution. With the heat and drag terms from equations (5) and (8), we can calculate the thermal evolution of the DM and baryon fluids as

$$\dot{T}_b = -2HT_b + 2\dot{Q}_b/3 + \Gamma_C(T_\gamma - T_b) \quad (10a)$$

$$\dot{T}_\chi = -2HT_\chi + 2\dot{Q}_\chi/3 \quad (10b)$$

$$\dot{x}_e = -C[n_H \mathcal{A}_B x_e^2 - 4(1 - x_e) \mathcal{B}_B e^{3E_0/(4T_\gamma)}] \quad (10c)$$

$$\dot{v}_{\chi,b} = -Hv_{\chi,b} - D(v_{\chi,b}) \quad (10d)$$

where H is the Hubble parameter at time t , C is the recombination factor^{35–37}, E_0 is the ground-level energy of hydrogen, and \mathcal{A}_B and \mathcal{B}_B are the effective case B recombination/reionization coefficients²³. We have ignored photoheating and recombination cooling, as well as possible baryonic heating due to DM annihilations, if these were present³⁸. Here T_γ is the temperature of the CMB photons, and the Compton thermalization rate is

$$\Gamma_C = \frac{8\sigma_T a_T T_\gamma^4 x_e}{3(1 + f_{\text{He}}) m_e c} \quad (11)$$

where σ_T is the Thomson cross-section, and a_T is the Stefan–Boltzmann constant^{36,39}.

Stellar production. If the mini-charged particles were lighter than about 100 keV, they could be produced in stars, such as white dwarfs and red giants, cooling these objects too rapidly²⁴. This tightly constrains their charge, although given that BBN already rules out masses above that limit we do not show these constraints in Fig. 1.

N_{eff} constraint. We can estimate for what value of the mini-charge dark photons would be produced, by requiring that the timescale for two mini-charged particles to annihilate into dark photons is longer than the Hubble time. For mini-charged particles in thermal equilibrium with standard-model particles in the early Universe, their rate of annihilation into dark photons is²⁹

$$\Gamma_{\chi\bar{\chi} \rightarrow \gamma'\gamma'} = n_\chi \sigma v \approx 10^{-3} g'^4 T_\gamma \quad (12)$$

where g' is the coupling constant between χ and γ' . By requiring this rate to be smaller than $H \approx T^2/M_{\text{pl}}$, where M_{pl} is the reduced Planck mass, we can obtain a constraint on g' , so that DM does not annihilate to dark photons before $T_\gamma = m_\chi$. Because the DM mini-charge is the product of the dark-photon mixing κ and the dark coupling g' , and we require $\kappa < 1$, this translates into a constraint

$$\epsilon \lesssim 2 \times 10^{-5} \left(\frac{m_\chi}{\text{MeV}} \right)^{1/4} \quad (13)$$

for $m_\chi \leq \Lambda_{\text{QCD}} \approx 200 \text{ MeV}$, where Λ_{QCD} is the quantum chromodynamics scale, which we label as N_{eff} in Fig. 1. Annihilations of χ particles into $\gamma\gamma'$, or Compton-like processes ($\chi\gamma \rightarrow \chi\gamma'$) would be suppressed by a κ^2 factor, and at most change the constraint by a $\mathcal{O}(1)$ factor. Notice that, even for $\kappa \approx 1$, if the dark photon is massless we can define the photon to be the linear combination of bosons that couples to our sector, in which case only mini-charged particles would interact with baryons. Here we have assumed that χ is a spin-1/2 particle, and we note that this constraint can, of course, be tightened if $\kappa \ll 1$, and can extend to DM masses as high as 1 GeV (ref.²⁹). We note, however, that even though dark photons are a possible way to obtain mini-charged particles⁴⁰, they might not be necessary⁴¹, which would render these constraints invalid.

Thermal-relic mini-charge. To compute the mini-charge required to produce the right DM abundance, we use the approximate formula $\Omega h^2 \approx 0.1 (x_f/10) [10^{-26} \text{ cm}^3 \text{ s}^{-1}/(\sigma v)]$, where $x_f \equiv m_\chi/T_f$ is the freeze-out temperature, and for mini-charged DM the annihilation cross-section to fermions is²⁰

$$(\sigma v) = \frac{\pi \alpha^2 \epsilon^2}{m_\chi^2} \sqrt{1 - \frac{m_f^2}{m_\chi^2}} \left(1 + \frac{m_f^2}{2m_\chi^2} \right) \quad (14)$$

We will ignore any dark-sector interactions and for simplicity consider only annihilation into electron–positron pairs. To obtain a simple estimate, we further approximate x_f to be a constant, as it only depends logarithmically on the DM mass and charge, and find the region of the $\epsilon - m_\chi$ plane that produces the right DM relic abundance. In the thermal-relic calculation we have assumed $f_{\text{dm}} = 1$, although of course to obtain a small fraction of DM with mini-charges, the rest of it would have to form dark atoms, or otherwise have a larger mini-charge.

Cosmology of mini-charged particles. Only a sliver of parameter space is compatible with the right relic cross-section, so one can wonder how mini-charged particles were produced. If they were heavier than about 200 MeV they could annihilate to light dark-sector particles, while not leaving a trace in current CMB probes⁴². However, for DM lighter than about 200 MeV, the standard-model plasma has been heated by the QCD phase transition, and any populated light degrees of freedom in the dark sector would alter the CMB anisotropies, which is strongly disfavoured by Planck data¹. In that case, other mechanisms would have to be invoked to set the right DM abundance⁴³.

Direct-detection constraints. Interestingly, mini-charged particles can remain in the Galactic disk if they cool efficiently, which for DM–electron cooling requires¹⁶ $\epsilon > 10^{-5}(m_\chi/\text{MeV})^{1/2}$, or otherwise can be achieved through dark-sector interactions. Moreover, we estimate the Galactic magnetic-field energy density to be at least three orders of magnitude smaller than the DM kinetic energy density in the solar vicinity. Thus, DM could be able to breach through magnetic field lines and re-enter the disk (albeit altering the magnetic field structure of our Galaxy). We note, however, that this might not lead to direct-detection signals in dark-matter detectors, as the gyroradius of one of these particles would be $r_g \approx 10 \text{ km} (m_\chi/\text{MeV})(\epsilon/10^{-5})^{-1}$ on the terrestrial magnetic field of $B_\oplus \approx 0.1 \text{ G}$. This would imply, though, that these particles can produce atmospheric ionizations, acting as a *nox borealis*, similar to the regular aurora borealis produced by solar-wind particles. We point out, however, that Earth-based experiments could be sensitive to even a minuscule trace of mini-charged particles, as these can interact rather strongly. Given that neither disk ejection, owing to the complex astrophysics of the interstellar medium, nor the terrestrial magnetic field would have perfect efficiency in shielding the Earth from these particles, there might be hope for direct detection, especially through space-based experiments. An example is the limits from the X-ray calorimeter⁴⁴, which, however, do not constrain these particles if their masses are below about 100 MeV. Additionally, we estimate that torsion-balance experiments⁴⁵, which can constrain accelerations as small as $10^{-13} \text{ cm s}^{-2}$, could be sensitive to mini-charges of the order of $10^{-6}(m_\chi/\text{MeV})$, if the DM density on the Earth's surface was 1% of its usual value. This result is comparable to that required for baryonic cooling during cosmic dawn, although the specific number is controlled by the fraction of the DM that diffuses to the solar vicinity. More importantly, the cross-section of these mini-charged particles would be similar to the atmospheric column density, so any constraints obtained on the surface would depend strongly on the DM momentum loss during atmospheric entry.

The 21-cm temperature. The brightness temperature of the 21-cm line can be written as⁴⁶

$$T^{(21)} = 27\text{mK} \left(\frac{T_s - T_\gamma}{T_s} \right) \left(\frac{x_{\text{HI}} \Omega_b h^2}{\Omega_m h^2} \right) \left(\frac{0.15}{0.023} \frac{1+z}{10} \right)^{1/2} \quad (15)$$

where x_{HI} is the neutral-hydrogen fraction, h is the reduced Hubble constant, $\Omega_m = \Omega_c + \Omega_b$, and T_s is the spin temperature of the hydrogen gas. We use the solutions for T_b of equations (10a–d), assuming full Lyman- α coupling (so $T_s = T_b$)⁴⁷, to obtain the sky-averaged 21-cm temperature^{48,49}

$$T_{\text{avg}}^{(21)} \equiv \langle T^{(21)} \rangle = \int d\nu_{\chi,b} \mathcal{P}(\nu_{\chi,b}) T^{(21)} [T_b(\nu_{\chi,b})] \quad (16)$$

Interactions with the neutral medium. As a check, we have estimated the interactions of mini-charged particles with the neutral baryonic medium through Linhard's formula^{20,50}, and found that they are always subdominant, by at least four orders of magnitude.

Other models. In this work we have assumed that mini-charged particles interact through a massless dark photon. However, our results apply to any dark photon lighter than the typical momentum transfer, which is between about 1 eV and 1 keV for DM masses between 1 MeV and 1 GeV. Additionally, we can easily translate our results for a DM mini-charge to a new hydrophilic or leptophilic DM–baryon interaction mediated by a light scalar ϕ . For $f_{\text{dm}} = 1$, we found that a DM mini-charge of $\epsilon \approx 10^{-8} \times (m_\chi/\text{MeV})$ is sufficient to decrease the baryonic

temperature by a factor of 2 (although we remind the reader that this case is ruled out for mini-charges). Even ignoring DM self interactions⁵¹, and setting the ϕ -DM coupling g_χ to unity, the ϕ -nucleon coupling required is $g_N = 8\pi\alpha\epsilon(\bar{x}_e)^{-1/2} \approx 2 \times 10^{-11} \times (m_\chi/\text{MeV})$, where $\bar{x}_e \approx 2 \times 10^{-4}$ during the era of interest. Similarly, the ϕ -electron coupling required would be $g_e = 8\pi\alpha\epsilon(\bar{x}_e m_e/m_p)^{-1/2} \approx 10^{-9} \times (m_\chi/\text{MeV})$. For DM in the MeV to GeV range (and thus mediators with $m_\phi \lesssim \text{keV}$), these couplings are constrained by stellar cooling, whereas for lighter DM the mediator would give rise to an anomalous fifth force^{52,53}.

Dark-matter self-thermalization. In this work, we have conservatively assumed that the small fraction of DM that is charged does not thermalize with the rest of the DM. If it did, one could simply rescale our results for ϵ from the $f_{\text{dm}} = 1$ case by $(f_{\text{dm}})^{-1/2}$.

Code availability. The code used to generate the thermodynamical evolution in equations (10a–d) is available upon request.

Data availability. The datum of 21-cm temperature at z_{central} used in Fig. 2 was obtained from ref.¹⁰.

31. Clarke, T. E., Kronberg, P. P. & Boehringer, H. A new radio-X-ray probe of galaxy cluster magnetic fields. *Astrophys. J.* **547**, L111–L114 (2001).
32. Heikinheimo, M., Raidal, M., Spethmann, C. & Veermäe, H. Dark matter self-interactions via collisionless shocks in cluster mergers. *Phys. Lett. B* **749**, 236–241 (2015).
33. Kadota, K., Sekiguchi, T. & Tashiro, H. A new constraint on millicharged dark matter from galaxy clusters. Preprint at <https://arxiv.org/abs/1602.04009> (2016).
34. Tashiro, H., Kadota, K. & Silk, J. Effects of dark matter–baryon scattering on redshifted 21 cm signals. *Phys. Rev. D* **90**, 083522 (2014).
35. Peebles, P. J. E. Recombination of the primeval plasma. *Astrophys. J.* **153**, 1 (1968).
36. Ali-Haïmoud, Y. & Hirata, C. M. Ultrafast effective multi-level atom method for primordial hydrogen recombination. *Phys. Rev. D* **82**, 063521 (2010).
37. Chluba, J. & Thomas, R. M. Towards a complete treatment of the cosmological recombination problem. *Mon. Not. R. Astron. Soc.* **412**, 748–764 (2011).
38. Slatyer, T. R. Energy injection and absorption in the cosmic dark ages. *Phys. Rev. D* **87**, 123513 (2013).
39. Ma, C.-P. & Bertschinger, E. Cosmological perturbation theory in the synchronous and conformal Newtonian gauges. *Astrophys. J.* **455**, 7–25 (1995).
40. Holdom, B. & Two, U. 1's and epsilon charge shifts. *Phys. Lett. B* **166**, 196–198 (1986).
41. Batell, B. & Gherghetta, T. Localized U(1) gauge fields, millicharged particles, and holography. *Phys. Rev. D* **73**, 045016 (2006).
42. Brust, C., Kaplan, D. E. & Walters, M. T. New light species and the CMB. *J. High Energy Phys.* **12**, 058 (2013).
43. D'Agnolo, R. T. & Ruderman, J. T. Light dark matter from forbidden channels. *Phys. Rev. Lett.* **115**, 061301 (2015).
44. Erickcek, A. L., Steinhardt, P. J., McCommon, D. & McGuire, P. C. Constraints on the interactions between dark matter and baryons from the X-ray Quantum Calorimetry Experiment. *Phys. Rev. D* **76**, 042007 (2007).
45. Wagner, T. A., Schlamminger, S., Gundlach, J. H. & Adelberger, E. G. Torsion-balance tests of the weak equivalence principle. *Class. Quantum Gravity* **29**, 184002 (2012).
46. Pritchard, J. R. & Loeb, A. Evolution of the 21 cm signal throughout cosmic history. *Phys. Rev. D* **78**, 103511 (2008).
47. Madau, P., Meiksin, A. & Rees, M. J. 21-cm tomography of the intergalactic medium at high redshift. *Astrophys. J.* **475**, 429 (1997).
48. Pritchard, J. R. & Loeb, A. Constraining the unexplored period between the dark ages and reionization with observations of the global 21 cm signal. *Phys. Rev. D* **82**, 023006 (2010).
49. Cohen, A., Fialkov, A. & Barkana, R. Charting the parameter space of the 21-cm power spectrum. Preprint at <https://arxiv.org/abs/1709.02122> (2017).
50. Lindhard, J. & Scharff, M. Energy dissipation by ions in the keV region. *Phys. Rev.* **124**, 128–130 (1961).
51. Harvey, D., Massey, R., Kitching, T., Taylor, A. & Tittley, E. The non-gravitational interactions of dark matter in colliding galaxy clusters. *Science* **347**, 1462–1465 (2015).
52. Knapen, S., Lin, T. & Zurek, K. M. Light dark matter: models and constraints. *Phys. Rev. D* **96**, 115021 (2017).
53. Stubbs, C. W. et al. Search for an intermediate-range interaction. *Phys. Rev. Lett.* **58**, 1070–1073 (1987).

An increase in the $^{12}\text{C} + ^{12}\text{C}$ fusion rate from resonances at astrophysical energies

A. Tumino^{1,2*}, C. Spitaleri^{2,3}, M. La Cognata², S. Cherubini^{2,3}, G. L. Guardo^{2,4}, M. Gulino^{1,2}, S. Hayakawa^{2,5}, I. Indelicato², L. Lamia^{2,3}, H. Petruscu⁴, R. G. Pizzone², S. M. R. Puglia², G. G. Rapisarda², S. Romano^{2,3}, M. L. Sergi², R. Spartá² & L. Trache⁴

Carbon burning powers scenarios that influence the fate of stars, such as the late evolutionary stages of massive stars¹ (exceeding eight solar masses) and superbursts from accreting neutron stars^{2,3}. It proceeds through the $^{12}\text{C} + ^{12}\text{C}$ fusion reactions that produce an alpha particle and neon-20 or a proton and sodium-23—that is, $^{12}\text{C}(^{12}\text{C}, \alpha)^{20}\text{Ne}$ and $^{12}\text{C}(^{12}\text{C}, p)^{23}\text{Na}$ —at temperatures greater than 0.4×10^9 kelvin, corresponding to astrophysical energies exceeding a megaelectronvolt, at which such nuclear reactions are more likely to occur in stars. The cross-sections⁴ for those carbon fusion reactions (probabilities that are required to calculate the rate of the reactions) have hitherto not been measured at the Gamow peaks⁴ below 2 megaelectronvolts because of exponential suppression arising from the Coulomb barrier. The reference rate⁵ at temperatures below 1.2×10^9 kelvin relies on extrapolations that ignore the effects of possible low-lying resonances. Here we report the measurement of the $^{12}\text{C}(^{12}\text{C}, \alpha_{0,1})^{20}\text{Ne}$ and $^{12}\text{C}(^{12}\text{C}, p_{0,1})^{23}\text{Na}$ reaction rates (where the subscripts 0 and 1 stand for the ground and first excited states of ^{20}Ne and ^{23}Na , respectively) at centre-of-mass energies from 2.7 to 0.8 megaelectronvolts using the Trojan Horse method^{6,7} and the deuteron in ^{14}N . The cross-sections deduced exhibit several resonances that are responsible for very large increases of the reaction rate at relevant temperatures. In particular, around 5×10^8 kelvin, the reaction rate is boosted to more than 25 times larger than the reference value⁵. This finding may have implications such as lowering the temperatures and densities⁸ required for the ignition of carbon burning in massive stars and decreasing the superburst ignition depth in accreting neutron stars to reconcile observations with theoretical models³.

We measured the $^{12}\text{C}(^{14}\text{N}, \alpha^{20}\text{Ne})^2\text{H}$ and $^{12}\text{C}(^{14}\text{N}, p^{23}\text{Na})^2\text{H}$ three-body processes in the quasi-free kinematic regime using the Trojan Horse Method (THM). The THM is an indirect technique with which to measure low-energy nuclear reactions unhindered by the Coulomb barrier and free of electron screening. The experimental and analysis procedures are detailed in Methods sections ‘THM basic features’, ‘One-level many-channel THM formalism’, ‘Experimental setup and channel selection’ and ‘Deuteron momentum distribution’. The experiment was performed at INFN, Laboratori Nazionali del Sud, Italy. A 30-MeV ^{14}N beam accelerated by the MP Tandem accelerator was delivered onto a carbon target. The detection setup consisted of two silicon telescopes, devoted to the detection of α -d and p -d coincidences. The occurrence and the dominance of the quasi-free mechanism⁵ was indicated by the agreement between the shapes of the experimental and the theoretical d momentum distributions (Extended Data Fig. 1).

The THM experimental yields projected onto the $^{12}\text{C}-^{12}\text{C}$ relative energy variable, the centre-of-mass energy E_{cm} , are shown as black dots in Fig. 1a ($^{20}\text{Ne} + \alpha_0$), Fig. 1b ($^{20}\text{Ne} + \alpha_1$), Fig. 1c ($^{23}\text{Na} + p_0$) and Fig. 1d ($^{23}\text{Na} + p_1$). A smooth four-body background due to $^{16}\text{O} + \alpha + \alpha + d$ was subtracted from the THM yields for the $^{20}\text{Ne} + \alpha_{0,1}$ channels. Error bars display the statistical errors and account for background subtraction uncertainty, when applicable, combined in quadrature.

A modified one-level many-channel R -matrix analysis was carried out including the excited states of the ^{24}Mg nucleus reported in Extended Data Table 1^{9–13}. The fraction of the total fusion yield from α and p channels^{14,15} other than $\alpha_{0,1}$ and $p_{0,1}$ was neglected with estimated errors at $E_{\text{cm}} < 2$ MeV lower than 1% and 2% for the α and p channels, respectively (see Methods section ‘Modified R -matrix analysis’).

The results are shown in Fig. 1a–d as red lines and with light-red shading indicating the uncertainties on the resonance parameters, including correlations. Agreement with experimental data is fair and confirmed by the reduced χ^2 (that is, $\bar{\chi}^2$) values of 0.73 for $^{20}\text{Ne} + \alpha_0$, 1.06 for $^{20}\text{Ne} + \alpha_1$, 0.54 for $^{23}\text{Na} + p_0$ and 1.34 for $^{23}\text{Na} + p_1$. The resonance structure observed in the excitation functions is consistent with ^{24}Mg level energies reported in the literature, with some tendency for the even- J states to be clustered¹¹ at about 1.5 MeV. The THM-reduced widths thus entered a standard R -matrix code¹⁶ and the $S(E)$ factors (see Methods section ‘Astrophysical $S(E)$ factor’) for the four reaction channels were determined.

The results are shown in Fig. 2a ($^{20}\text{Ne} + \alpha_0$), Fig. 2b ($^{20}\text{Ne} + \alpha_1$), Fig. 2c ($^{23}\text{Na} + p_0$) and Fig. 2d ($^{23}\text{Na} + p_1$), in terms of the modified $S(E)$ factor^{15,17}, $S(E)^*$, (see Methods section ‘Astrophysical $S(E)$ factor’). The black line and grey shading in each panel represent the best-fit curve and the range defined by the total uncertainties, respectively. The grey shading is the result of R -matrix calculations with lower and upper values of the resonance parameters provided by their errors after being combined with the normalization one. Excursions from the midline range from 11% to 20%.

The resonant structures are superimposed onto a flat nonresonant background¹⁵ of 0.4×10^{16} MeV b. Unitarity of the S matrix is guaranteed within the experimental uncertainties. Normalization to direct data was done in the E_{cm} window 2.50–2.63 MeV of the $^{20}\text{Ne} + \alpha_1$ channel, where a sharp resonance corresponding to the 16.5-MeV level⁹ of ^{24}Mg appears and available data^{15,18–20} in this region are the most accurate of those overlapping with THM data. By scaling to the resonance by means of a weighted normalization, the resulting normalization error is 5%, shown as grey shading in Fig. 2a–d, combined in quadrature with errors on the resonance parameters.

Existing direct data below $E_{\text{cm}} = 3$ MeV are shown as red filled circles¹⁵, purple filled squares¹⁸, blue empty diamonds¹⁹, blue filled stars²⁰ and green filled triangles²¹ in Fig. 2. Their low-energy limit is mostly fixed by the background due to hydrogen contamination in the targets^{18–21} and the higher $S(E)$ values for the p_1 channel in some of them^{19–21} were attributed to Coulomb excitation of ^{23}Na contamination in the targets or collimators^{15,20}. Disregarding these cases, agreement between THM results and direct data are apparent within the experimental errors, except for the direct low-energy limit around 2.14 MeV, where THM data do not confirm the claim of a strong resonance; instead, there is a nearby resonance at 2.095 MeV, about one order of magnitude less intense in the $^{20}\text{Ne} + \alpha_1$ channel (see Fig. 2b) and with similar intensity in the $^{23}\text{Na} + p_1$ one (see Fig. 2d). The present

¹Facoltà di Ingegneria e Architettura, Università degli Studi di Enna “Kore”, Enna, Italy. ²INFN, Laboratori Nazionali del Sud, Catania, Italy. ³Dipartimento di Fisica e Astronomia, Università degli Studi di Catania, Catania, Italy. ⁴Horia Hulubei National Institute for R&D in Physics and Nuclear Engineering, Bucharest-Magurele, Romania. ⁵Center for Nuclear Studies, The University of Tokyo, Tokyo, Japan. *e-mail: tumino@lns.infn.it

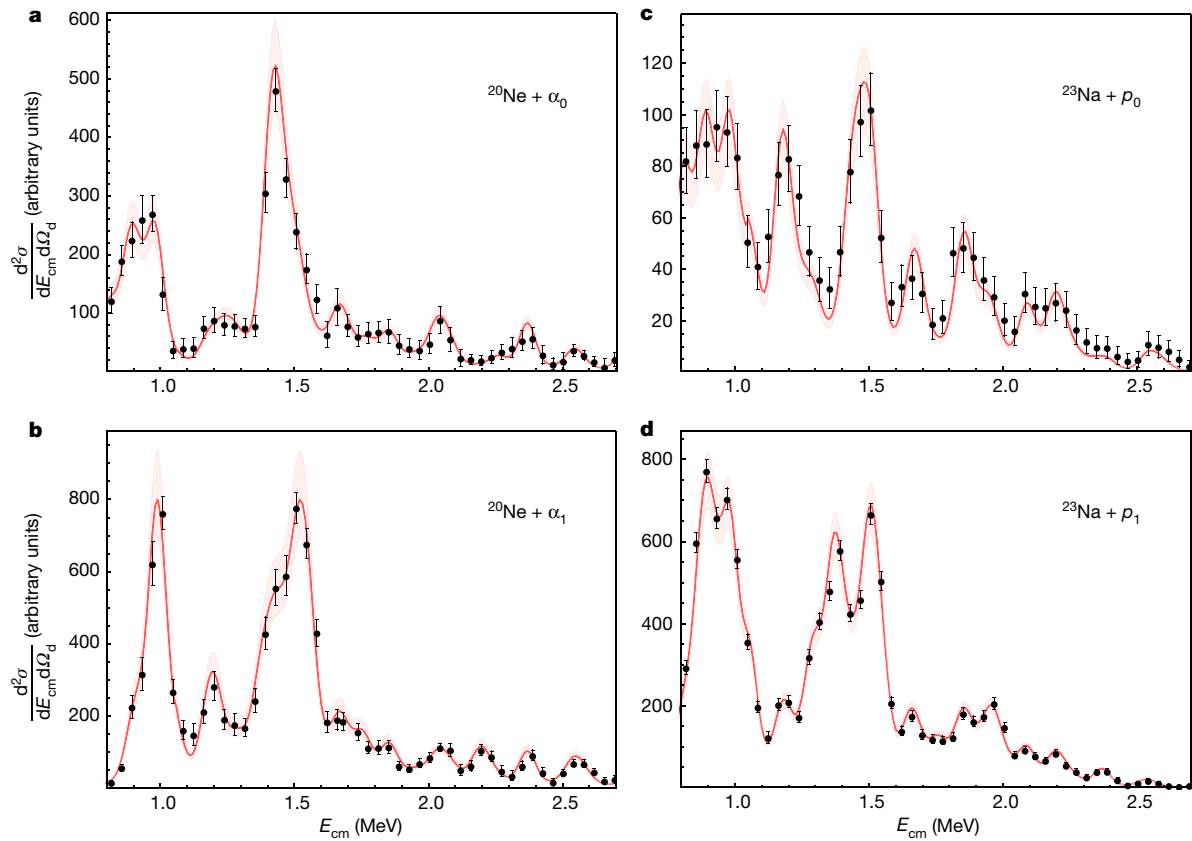


Fig. 1 | Excitation functions from THM experimental yields. The quasi-free cross-section for the four channels $^{20}\text{Ne} + \alpha_0$ (a), $^{20}\text{Ne} + \alpha_1$ (b), $^{23}\text{Na} + p_0$ (c) and $^{23}\text{Na} + p_1$ (d) is projected onto the E_{cm} variable (black dots). Error bars denote $\pm 1\sigma$ uncertainties and account for background

subtraction (combined in quadrature). Red lines and light-red shading represent the results of the modified R -matrix fits and the related uncertainties, respectively.

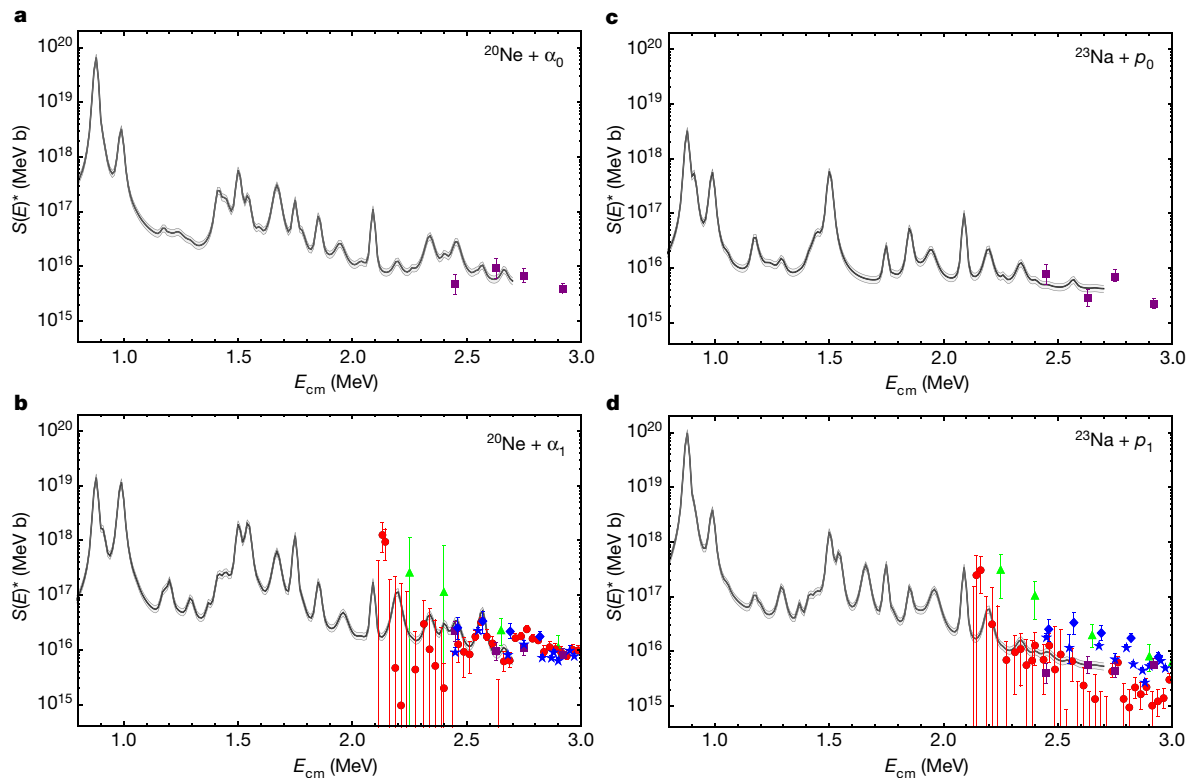


Fig. 2 | $^{12}\text{C} + ^{12}\text{C}$ astrophysical $S(E)^*$ factors. The THM $S(E)^*$ factors for the four channels $^{20}\text{Ne} + \alpha_0$ (a), $^{20}\text{Ne} + \alpha_1$ (b), $^{23}\text{Na} + p_0$ (c) and $^{23}\text{Na} + p_1$ (d) are shown as black lines. The available direct data in the E_{cm} range investigated are reported as red filled circles¹⁴, purple filled squares¹⁷,

blue empty diamonds¹⁸, blue filled stars¹⁹ and green filled triangles²⁰. The upper and lower grey lines mark the range arising from $\pm 1\sigma$ uncertainties on resonance parameters plus the normalization to direct data in the $^{20}\text{Ne} + \alpha_1$ channel at $E_{\text{cm}} = 2.50\text{--}2.63$ MeV.

result is in agreement with spectroscopy studies^{9,22} that report a dip at 2.14 MeV and no particularly strong α state at around 2.1 MeV. Further agreement is found with unpublished experimental data down to $E_{\text{cm}} = 2.15$ MeV for the $^{12}\text{C}(^{12}\text{C}, p_{0,1})^{23}\text{N}$ reactions²³. Our result is also consistent within experimental errors with the total $S(E)^*$ from a recent experiment at higher energies²⁴, which was calculated at the overlapping $E_{\text{cm}} = 2.68 \pm 0.08$ MeV.

The reaction rates for the four processes were calculated from the THM $S(E)^*$ factors using the standard formula⁴ and summed to obtain the total $^{12}\text{C} + ^{12}\text{C}$ reaction rate. Its numerical values are given in Extended Data Table 2 (see Methods section ‘Numerical values of the $^{12}\text{C} + ^{12}\text{C}$ reaction rate’). We recommend an analytical expression for the reaction rate and for its upper and lower limits, based on the same formulae as reported in the REACLIB library²⁵. This expression is valid in the temperature range $0.1 \text{ GK} \leq T \leq 3 \text{ GK}$ with an accuracy better than 0.7% ($\chi^2 = 0.1$), which refers to the maximum difference between the analytical function and the centroids of the experimental points. This is given by:

$$N_A \langle \sigma v \rangle = \sum_{i=1}^3 f_i = \sum_{i=1}^3 \exp[a_{i1} + a_{i2}T^{-1} + a_{i3}T^{-1/3} + a_{i4}T^{1/3} + a_{i5}T + a_{i6}T^{5/3} + a_{i7}\ln(T)] \quad (1)$$

Parameters a_{ij} with $1 < i < 3$ and $1 < j < 7$ are given in Table 1, with subscripts ‘u’ and ‘l’ for the upper and lower limits. They result from a fit performed using the NUCASTRODATA toolkit (<http://www.nucastrodata.org/>).

The total THM reaction rate was divided by the reference rate⁵. The resulting ratio is shown in Fig. 3. The black line represents the rate from the present work, with the grey shading defining the region fixed by the total uncertainty (Methods section ‘Numerical values of the $^{12}\text{C} + ^{12}\text{C}$ reaction rate’), whereas the red line refers to the reference rate⁵.

The light-blue shading shows the temperature range relevant for superbursts (about 0.4–0.5 GK), the light-red shading highlights typical temperatures for hydrostatic carbon burning in massive stars (about 0.6–1.0 GK in the core and up to 1.2 GK in the shell, depending on the stellar mass), whereas the light-green shading marks the temperatures of explosive carbon burning (about 1.8–2.5 GK). As shown in Fig. 3, the reaction rate changes below 2 GK with an increase with respect to the reference non-resonant one⁵ from a factor of 1.18 at 1.2 GK ($^{***}P < 0.001$) to a factor of more than 25 at 0.5 GK ($^{****}P < 0.00001$). The latter increase, mainly due to the resonances around $E_{\text{cm}} = 1.5$ MeV, supports the conjectured fiducial value³ required to reduce the theoretical superburst ignition depths in accreting neutron stars by a factor of 2 for a range of realistic parameters and core neutrino emissivities. This change matches the observationally inferred ignition depths and can be translated into an ignition temperature below 0.5 GK, compatible with the calculated crust temperature. In other words, carbon burning can trigger superbursts. A similar decrease in temperature is obtained by using the crust Urca shell neutrino emissivities²⁶, recently invoked to explain the cooling of the outer neutron star crust, while thermally decoupling the surface layers from the deeper crust. Under this hypothesis, a revision of current superburst models and predicted light curves is required and our finding could represent the missing heat source in the standard carbon ignition scenario.

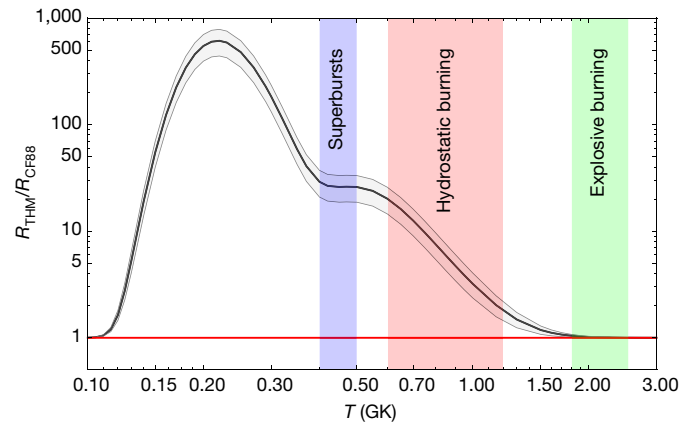


Fig. 3 | $^{12}\text{C} + ^{12}\text{C}$ reaction rate ratio. Ratio between the total THM $^{12}\text{C} + ^{12}\text{C}$ reaction rate (black line) and the reference one⁴ (red line). The grey shading defines the region spanned owing to the $\pm 1\sigma$ uncertainties. The coloured shading marks typical temperature regions for carbon burning in different scenarios: light blue for superbursts from accreting neutron stars, light red for hydrostatic carbon burning in massive stars and light green for explosive carbon burning; comparison with the red line (non-resonant assumption) gives $^{***}P < 0.001$ in the region of hydrostatic burning and $^{****}P < 0.00001$ at superburst temperatures.

In the hydrostatic carbon burning regime, the present rate change will lower the temperatures and densities at which ^{12}C ignites in massive post-main-sequence stars. We make use of stellar modelling⁸ for core carbon burning of a star of 25 solar masses to determine that the ignition temperature and density would decrease to 10% and 30% respectively. This would reduce the neutrino losses, thus causing the carbon burning stage to occur for a lifetime (of the carbon burning phase) longer by up to a factor of 70. The new rate would also affect abundances of species that are the main fuel for subsequent evolutionary phases. However, such abundances are influenced also by the ratio of the α to p yields if it deviates from unity. From the present experiment, the average value of this ratio is around 2. In particular, at 0.8 GK this ratio is 1.6 ± 0.4 , and it becomes 2.2 ± 0.6 at 2 GK. The $^{12}\text{C} + ^{12}\text{C}$ rate is also the most important nuclear physics input governing the minimum stellar mass M_{up} required for hydrostatic carbon burning to occur. M_{up} is fundamental to our understanding, for instance, of the evolution of supernova progenitors and the white dwarf luminosity functions. From the present result, we consider that the present value of M_{up} will not be strongly affected, in contrast to what has been predicted^{27,28} when assuming a much larger increase (up to nine orders of magnitude) in the reaction rate, but it is worth noticing that stellar models are also very sensitive to small changes of this parameter. However, a sound evaluation of M_{up} requires a better understanding of the ratio of the initial mass to the final core mass.

Below 0.4 GK the rate experiences a huge increase by up to a factor of 800 owing to the lowest-energy resonances occurring around $E_{\text{cm}} = 1$ MeV. It has been conjectured that the existence of such low-energy resonances might shift the ignition curve of type Ia supernovae to lower central densities³. This should be assessed for the various progenitor scenarios. Much additional work is needed

Table 1 | Coefficients of the analytical function of the $^{12}\text{C} + ^{12}\text{C}$ reaction rate using equation (1)

a_{ij}	f_1	f_2	f_3	f_{1u}	f_{2u}	f_{3u}	f_{1l}	f_{2l}	f_{3l}
a_{11}	1.22657×10^2	9.03221×10^1	2.28039×10^2	1.22687×10^2	9.03982×10^1	2.28056×10^2	3.21570×10^2	6.08741×10^2	3.14593×10^3
a_{12}	0.557112	−0.35888	−1.16039 × 10 ¹	0.557664	−0.35720	−1.15681 × 10 ¹	−0.815182	−1.42976 × 10 ¹	−2.26169 × 10 ¹
a_{13}	−905657 × 10 ¹	−6.17552 × 10 ¹	−2.40364 × 10 ²	−9.05616 × 10 ¹	−6.17282 × 10 ¹	−2.40343 × 10 ²	3.17671 × 10 ¹	3.43845 × 10 ²	1.36110 × 10 ³
a_{14}	−6.83561 × 10 ¹	−1.07514 × 10 ²	−9.21375 × 10 ¹	−6.83178 × 10 ¹	−1.07358 × 10 ²	−9.21156 × 10 ¹	−4.22173 × 10 ²	−1.11874 × 10 ³	−5.16494 × 10 ³
a_{15}	1.42906 × 10 ¹	7.20344 × 10 ¹	1.25411 × 10 ²	1.42891 × 10 ¹	7.20835 × 10 ¹	1.25484 × 10 ²	5.23691 × 10 ¹	1.73098 × 10 ²	7.85965 × 10 ²
a_{16}	−2.43583	−1.37501 × 10 ¹	−3.25984 × 10 ¹	−2.46506	−1.38060 × 10 ¹	−3.24417 × 10 ¹	−6.35869	−2.33743 × 10 ¹	−1.29447 × 10 ²
a_{17}	9.32623	−1.91793 × 10 ¹	−1.10903 × 10 ²	9.35304	−1.91920 × 10 ¹	−1.10961 × 10 ²	1.34509 × 10 ²	3.60334 × 10 ²	1.60224 × 10 ³

Coefficients of the analytical function (equation (1)) of the $^{12}\text{C} + ^{12}\text{C}$ reaction rate and of its upper and lower limits. They result from a fit of the numerical values given in Extended Data Table 2 using the reaction rate parameterizer from the NUCASTRODATA toolkit (<http://www.nucastrodata.org/>).

to determine the impact that the new $^{12}\text{C} + ^{12}\text{C}$ reaction rate will have in various astrophysical contexts.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0149-4>.

Received: 13 November 2017; Accepted: 28 February 2018;

Published online 23 May 2018.

1. Woosley, S. E., Heger, A. & Weaver, T. A. The evolution and explosion of massive stars. *Rev. Mod. Phys.* **74**, 1015–1071 (2002).
2. Keek, L. et al. First superburst from a classical low-mass X-ray binary transient. *Astron. Astrophys.* **479**, 177–188 (2008).
3. Cooper, R. L. et al. Possible resonances in the $^{12}\text{C} + ^{12}\text{C}$ fusion rate and superburst ignition. *Astrophys. J.* **702**, 660–671 (2009).
4. Iliadis, C. *Nuclear Physics of Stars*. (Wiley, Weinheim, 2007).
5. Caughlan, G. R. & Fowler, W. A. Thermonuclear reaction rates V. *At. Data Nucl. Data Tables* **40**, 283–334 (1988).
6. Spitaleri, C. et al. The Trojan Horse Method in nuclear astrophysics. *Phys. At. Nucl.* **74**, 1725–1739 (2011).
7. Tribble, R. et al. Indirect techniques in nuclear astrophysics: a review. *Rep. Prog. Phys.* **77**, 106901–106950 (2014).
8. Pignatari, M. et al. The $^{12}\text{C} + ^{12}\text{C}$ reaction and the impact on nucleosynthesis in massive stars. *Astrophys. J.* **762**, 31–54 (2013).
9. Abegg, R. & Davis, C. A. ^{12}Mg states observed via $^{20}\text{Ne}(\alpha, \alpha_0)^{20}\text{Ne}$. *Phys. Rev. C* **43**, 2523–2540 (1991).
10. Itoh, K. et al. Electroexcitation of giant multipole resonances in ^{24}Mg . *Phys. Rev. C* **23**, 945–959 (1981).
11. Yang, G. C. et al. Isoscalar multipole strength in ^{24}Mg through inelastic α scattering. *Phys. Rev. C* **13**, 1376–1387 (1976).
12. Vanhoy, J. R. et al. Proton resonances in ^{24}Mg from $E_x = 12.7$ to 15.7 MeV. *Phys. Rev. C* **36**, 920–932 (1987).
13. Bertrand, F. E. et al. Giant quadrupole resonance in $^{24,26}\text{Mg}$: a comparison of inelastic-scattering and α -capture experiments. *Phys. Rev. Lett.* **40**, 635–638 (1978).
14. Becker, H. W., Kettner, K. U., Rolfs, C. & Trautvetter, H. P. The $^{12}\text{C} + ^{12}\text{C}$ reaction at sub-Coulomb energies (II). *Z. Phys. A* **303**, 305–312 (1981).
15. Spillane, T. et al. $^{12}\text{C} + ^{12}\text{C}$ fusion reactions near the Gamow Energy. *Phys. Rev. Lett.* **98**, 122501–122505 (2007).
16. Lane, A. M. & Thomas, R. G. R -matrix theory of nuclear reactions. *Rev. Mod. Phys.* **30**, 257–353 (1958).
17. Aguilera, E. F. et al. New γ -ray measurements for $^{12}\text{C} + ^{12}\text{C}$ sub-Coulomb fusion: toward data unification. *Phys. Rev. C* **73**, 064601–064612 (2006).
18. Mazarakis, M. G. & Stephens, W. E. Experimental measurements of the $^{12}\text{C} + ^{12}\text{C}$ nuclear reactions at low energies. *Phys. Rev. C* **7**, 1280–1287 (1973).
19. High, M. D. & Cujec, B. The $^{12}\text{C} + ^{12}\text{C}$ sub-Coulomb fusion cross section. *Nucl. Phys. A* **282**, 181–188 (1977).
20. Kettner, K. U., Lorenz-Wirzba, H. & Rolfs, C. The $^{12}\text{C} + ^{12}\text{C}$ reaction at subcoulomb energies. *Z. Phys. A* **298**, 65–75 (1980).
21. Barrón-Palos, L. et al. Absolute cross sections measurement for the $^{12}\text{C} + ^{12}\text{C}$ system at astrophysically relevant energies. *Nucl. Phys. A* **779**, 318–332 (2006).
22. Cacioli, A. et al. Proton elastic scattering and proton induced γ -ray emission cross-sections on ^{23}Na from 2 to 5 MeV. *Nucl. Instrum. Meth. Phys. Res. B* **266**, 1392–1396 (2008).
23. Zickefoose, J. $^{12}\text{C} + ^{12}\text{C}$ Fusion: Measurement and Advances Toward the Gamow Energy. PhD thesis, Univ. of Connecticut (2011); <https://pqdtopen.proquest.com/doc/908637546.html?FMT=AI>.
24. Jiang, C. L. et al. Reaction rate for carbon burning in massive stars. *Phys. Rev. C* **97**, 012801–012806 (2018).
25. Thielemann, F.-K., Arnould, M. & Truran, J. W. Thermonuclear reactions rate from statistical model calculations. In *Advances in Nuclear Astrophysics (2nd IAP Workshop)* (ed. Vangioni-Flam, E.) 525 (Les Editions Frontières, Gif sur Yvette, 1987).
26. Schatz, H. et al. Strong neutrino cooling by cycles of electron capture and β decay in neutron star crust. *Nature* **505**, 62–65 (2014).
27. Bravo, E. et al. Type Ia supernovae and the $^{12}\text{C} + ^{12}\text{C}$ reaction rate. *Astron. Astrophys.* **535**, A114 (2011).
28. Straniero, O., Piersanti, L. & Cristallo, S. Do we really know Mup (i.e. the transition mass between type Ia and core-collapse supernova progenitors)? *J. Phys. Conf. Ser.* **665**, 012008 (2016).

Acknowledgements We thank V. Z. Goldberg for having inspired the idea of the experiment and for discussions and assistance. We thank A. M. Mukhamedzhanov for having developed the theoretical framework of the THM. The aid of the technical staff of INFN-LNS during the preparation of the experiment is gratefully acknowledged. We thank M. Wiescher and F. X. Timmes for comments.

Author contributions A.T. and C.S. proposed the experiment. A.T., C.S., M.L.C., G.L.G., I.L., L.L., R.G.P., S.M.R.P., R.S. and G.G.R. set up and ran the experiment, which lasted about one month. S.C., M.G., S.H., H.P., M.L.S., S.R. and L.T. participated in the data collection. A.T. performed the data reduction and analysis. M.L.C. developed the modified R -matrix code for the one-level many-channel case. A.T. and M.L.C. performed the statistical analysis. A.T. performed R -matrix calculations, interpreted the results, prepared the figures and wrote the manuscript. C.S. and M.L.C. contributed on the interpretation of the results. M.L.C. assisted with the figure preparation. L.L., R.G.P. and R.S. assisted with the astrophysical interpretation. A.T., C.S., M.L.C., S.C., G.L.G., I.L., L.L., R.G.P., G.G.R., R.S., S.R. and L.T. revised the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0149-4>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to A.T.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

THM basic features. The THM is an indirect technique aiming at measuring low-energy nuclear reactions unhindered by the Coulomb barrier and free of electron screening^{6,7,29}. It has been used to study several reactions related to fundamental astrophysical problems^{30–34}. In the THM, the low-energy cross-section of an A(x,b)B reaction is determined by selecting the quasi-free contribution of a suitable A(a,b)s reaction that is measured. In quasi-free kinematics, particle a, chosen for its xs cluster structure, is used to transfer the participant cluster x to induce the reaction with A, while the other constituent cluster s remains a spectator to the A(x,b)B sub-process⁶. Because the transferred nucleus x is virtual, its energy and momentum are not linked by the usual energy–momentum relation for a free particle. This gives the A(x,b)B reaction its half-off-the-energy-shell (HOES) character. The quasi-free A(a,b)s reaction can be sketched using a pole diagram (see Extended Data Fig. 2) with two vertices referring to a break-up (upper vertex) and to the A(x,b)B process (lower vertex). The A + a relative motion takes place at an energy above the Coulomb barrier, ensuring that the transfer of particle x occurs inside the nuclear field of A without undergoing Coulomb suppression or electron screening. However, the A + x reaction takes place at the sub-Coulomb relative energy E_{cm} because the excess of energy in the A + a relative motion is needed for the break-up of the Trojan Horse nucleus $a = (xs)$. From the principles of energy and momentum conservation, we obtain:

$$E_{\text{cm}} = \frac{m_x}{m_x + m_A} E_A - \frac{p_s^2}{2\mu_{\text{sf}}} + \frac{\mathbf{p}_s \cdot \mathbf{p}_A}{m_x + m_A} - B_{\text{xs}} \quad (1)$$

with m_i and \mathbf{p}_i the mass and momentum of particle i , $\mu_{ij} = m_i m_j / (m_i + m_j)$ the reduced mass of particles i and j , F the compound system ($F = A + x = b + B$) and $B_{\text{xs}} = m_s + m_x - m_a$ the binding energy of clusters x and s inside a . E_{cm} can vary within a range determined by the momentum of the spectator particle, \mathbf{p}_s , or its emission angle. As for \mathbf{p}_s , its values should not exceed the theoretical upper limit for the relative momentum p_{xs} between x and s (in the laboratory system, $\mathbf{p}_{\text{xs}} = \mathbf{p}_x = -\mathbf{p}_s$) represented by the on-the-energy-shell bound state wave number $\kappa_{\text{xs}} = (2\mu_{\text{xs}} B_{\text{xs}})^{1/2}$. This is the condition for the quasi-free mechanism to be dominant, for example, for the HOES cross-section to approach the on-energy-shell cross-section minimizing distortions. For the $^{14}\text{N} = (^{12}\text{Cd})$ system, $\kappa_{\text{xs}} = 181 \text{ MeV } c^{-1}$ (where c is the velocity of light), exceeding by far the experimental p_s upper limit of about $80 \text{ MeV } c^{-1}$, which is fixed by the phase space populated in the present experiment. In the plane-wave impulse approximation, the three-body cross-section can be factorized into two terms corresponding to the vertices of Extended Data Fig. 1 and given by:

$$\frac{d^3\sigma}{d\Omega_B d\Omega_b dE_B} \propto \text{KF} \left| \tilde{\Phi}(p_{\text{xs}}) \right|^2 \left[\frac{d^2\sigma_{\text{xA} \rightarrow \text{bB}}}{dE_{\text{xA}} d\Omega} \right]_{\text{HOES}} \quad (2)$$

where KF is a kinematical factor containing the final state phase space factor and it is a function of the masses, momenta and angles of the outgoing particles⁶; $|\tilde{\Phi}(p_{\text{xs}})|^2$ is the squared Fourier transform of the radial wave function for the $\chi(r_{\text{xs}})$ inter-cluster motion whose functional dependence is fixed by the xs system properties; $\left[\frac{d^2\sigma_{\text{xA} \rightarrow \text{bB}}}{dE_{\text{xA}} d\Omega} \right]_{\text{HOES}}$ is the HOES cross-section of the binary reaction.

One-level many-channel THM formalism. In the case of a multi-resonance A(x,b)B reaction, the so-called modified R -matrix approach has been developed^{35,36} to account for its HOES nature in the extraction of the reduced widths γ from the THM reaction yield. Because the transferred particle does not obey the mass–shell equation, no entrance-channel penetration factor is present, making it possible to reach astrophysical energies with no need of extrapolation. Yet the same reduced widths appear in the THM and in the on-energy-shell cross-sections, so the ones extracted from THM data can be used to determine the direct $S(E)$ factor, without HOES effects. For isolated non-interfering resonances, the one-level many-channel formula can be used, so that the THM A(x,b)B cross-section in the plane-wave impulse approximation^{35,37} takes the form:

$$\frac{d^2\sigma_{\text{xA} \rightarrow \text{c}'}}{dE_{\text{xA}} d\Omega_s} = \text{NF} \sum_i (2J_i + 1) \left| \frac{\sqrt{k_{\text{c}'}} \sqrt{2P_{\text{c}'}} M_i(p_{\text{xA}} R_{\text{xA}}) \gamma_{\text{xA}}^i \gamma_{\text{c}'}^i}{D_i(E_{\text{xA}})} \right|^2 \quad (3)$$

with NF a normalization factor; $k_{\text{c}'}$ and $P_{\text{c}'}$ are the exit-channel wave number and penetration factor (c' runs over all exit channels), E_{xA} and R_{xA} are the x-A entrance-channel relative energy and channel radius¹⁷ is set to 7.25 fm:

$$M_i(p_{\text{xA}} R_{\text{xA}}) = \left[(B_{\text{xA}} - 1) j_{l_i}(\rho) - \rho \frac{\partial j_{l_i}(\rho)}{\partial \rho} \right]_{\rho = p_{\text{xA}} R_{\text{xA}}} \quad (4)$$

where $j_{l_i}(\rho)$ is the spherical Bessel function for the l_i wave, $p_{\text{xA}} = \sqrt{2\mu_{\text{xA}} (E_{\text{xA}} + B_{\text{xs}})} / \hbar$, B_{xA} is an arbitrary boundary condition chosen to reproduce the observable resonance parameters^{38,39} and $D_i(E_{\text{xA}})$ is the R -matrix denominator of one-level multi-channel formulas¹⁵:

$$D_i(E_{\text{xA}}) = E_i - E_{\text{xA}} - \sum_c (\gamma_c^i)^2 (S_c - B_c) - I \sum_c 2P_c (\gamma_c^i)^2 \quad (5)$$

with the sum running over all the open channels c ; S_c and B_c are the shift function and the boundary condition for channel c and γ_c^i is the reduced width for the i th resonance and c channel, which enters the calculation of the on-energy-shell $S(E)$ factor free of electron screening and is not affected by the experimental energy resolution.

Experimental setup and channel selection. A ^{14}N beam at 30 MeV was delivered onto a carbon target, $100 \mu\text{g cm}^{-2}$ thick, with a spot size of 1 mm. The silicon telescopes were made up of a $38\text{-}\mu\text{m}$ ΔE -detector and a $1,000\text{-}\mu\text{m}$ position-sensitive E -detector (with intrinsic α resolution quoted as 0.3 mm for the position and about 0.5% for the energy) to measure the residual energy. They were placed symmetrically at either side of the beam direction, each covering laboratory angles 8° to 30° , and devoted to the detection of α -and-d and p -and-d coincidences. Angular conditions were selected to maximize the expected quasi-free contribution, fulfilling the requirement for the spectator particle d to retain its initial momentum inside ^{14}N . Channel selection was accomplished by gating on the ΔE - E two-dimensional plots to select coincident d and $\alpha(p)$ loci. A typical ΔE - E spectrum is shown in Extended Data Fig. 3, where p , d and α loci are clearly visible. Kinematics were reconstructed under the assumption of either a ^{20}Ne (for the $\alpha + d$ channel) or a ^{23}Na (for the $p + d$ channel) as an undetected particle. The Q -value variable was reported as a function of a kinematic variable, such as the energy or the angle of any one of the particles involved. In this representation, coincidence events of interest should lie on a horizontal line that cuts the Q -value axis at the expected value, because the Q -value depends only on the masses of the particles involved. A typical spectrum for the present experiment is shown in Extended Data Fig. 4 for the $^{12}\text{C}(^{14}\text{N}, \alpha^{20}\text{Ne})^2\text{H}$ reaction, where the Q -value is reported as a function of the α detection angle. Two dominant sharp horizontal loci appear, corresponding to the ground and first excited states of ^{20}Ne . They are highlighted by blue and red solid lines crossing the Q -value axis at -5.65 MeV and -7.28 MeV , respectively. This spectrum makes us confident of the quality of the calibration and of the correct selection of the reaction channel. Further data analysis was restricted to such events.

Deuteron momentum distribution. The d momentum distribution is a physical quantity very sensitive to the reaction mechanism. It keeps the same shape as inside ^{14}N only if the latter experiences quasi-free break-up. The agreement between the shapes of the theoretical and experimental momentum distributions is thus a compelling signature of occurrence of the quasi-free mechanism^{6,7}. To determine the d momentum distribution from the coincidence yield, the modulation due to possible contributions of ^{24}Mg states has to be removed. This is done over restricted ranges of E_{cm} and θ_{cm} of less than 30 keV and 5° , respectively. The kinematic factor KF, describing the phase space population, is divided out by performing a Monte Carlo simulation of the experimental setup with the angular ranges and detection thresholds of the experiment. The momentum distribution from the $^{12}\text{C}(^{14}\text{N}, \alpha^{20}\text{Ne})^2\text{H}$ reaction is shown as an example in Extended Data Fig. 1 as black filled circles. Data are projected in $8 \text{ MeV } c^{-1}$ bins over the momentum axis of the detected deuteron, p_d , with error bars including statistical errors only. The solid black line in the figure represents the theoretical behaviour normalized to experimental data. It is obtained from the Woods–Saxon ^{12}C d bound state potential with standard geometrical parameters $r_0 = 1.25 \text{ fm}$, $a = 0.65 \text{ fm}$ and $V_0 = 54.428 \text{ MeV}$, adjusted to give the experimental ground state $^{12}\text{C}_{\text{gs}}$ d binding energy in ^{14}N . A fair accordance ($\chi^2 = 0.2$) shows up, indicating that in the phase space region spanned in the present experiment the reaction mainly proceeds through a direct ^{12}C transfer. Thus, the plane-wave impulse approximation factorization of equation (2) can be relied on for the present investigation because no distortions are needed within experimental errors to describe the transfer process³⁷. This result agrees with previous work^{40,41} where a strong transfer component is found in similar kinematic conditions with d detected at forward angles. We remark that in the present experiment the d centre-of-mass angular range is about 11° – 50° and the coincidence mode triggers event acquisition. In those papers^{40,41}, it was taken into account that the transferred ^{12}C can be found also in its first excited 2^+ state at 4.44 MeV. From angular distribution analysis using a general expression for resonance reactions⁴², there is no evidence in our experimental data of a ^{12}C transfer in its first excited 2^+ state at 4.44 MeV. It turns out that only transfer of ^{12}C in its ground state is contributing. This result will be discussed in a future paper.

From the shape analysis of the momentum distribution, we could estimate the possible contribution of reaction mechanisms other than the quasi-free one to the extracted experimental yield. In particular, other contributing mechanisms, such as

'compound nucleus' or 'multistep transfer', are represented by an isotropic momentum distribution as a signature of loss of correlation in the deuteron momentum. Thus, a fit of the experimental shape of the momentum distribution was performed with a linear combination of the theoretical function for the quasi-free mechanism with a constant one, leaving coefficients as free parameters. The covariance matrix fit returns a contribution consistent with zero for the constant function within an uncertainty of 3% at a 2σ level, including correlations. This contribution to the overall uncertainty was neglected in the further extraction of cross-sections and reaction rates.

Modified R-matrix analysis. Level energies and J^π values were taken from the literature^{8,10–12}. In particular, J^π assignments were checked to be in agreement for the most prominent peaks in the present experiment through angular distribution analysis. Because the widths of the ^{24}Mg states at the relevant excitation energies are smaller than their average spacing^{8,43} (for example, at $E_{\text{cm}} \approx 1.5$ MeV, $\Gamma/D < 0.4$, with Γ and D the level widths and spacings, respectively.) and most importantly, owing to the smoothing effect of our experimental resolution, interference between them was not taken into account. Indeed, the energy resolution removes the problem of interference, making our result insensitive to it⁴⁴. Thanks to the non-overlapping nature of the levels involved, integration of the $d^3\sigma/dE_{\text{cm}}d\Omega_d d\Omega_{\text{cm}}$ over the $\alpha(p)$ emission angle in the centre-of-mass system of the $^{12}\text{C}(^{12}\text{C}, \alpha_0, \alpha_1)^{20}\text{Ne}$ and $^{12}\text{C}(^{12}\text{C}, p_{0,1})^{23}\text{N}$ sub-reactions could be easily performed. Because the experimental θ_{cm} range of the sub-reactions is about 140° to 180° , angular distributions outside this angular region were calculated by means of the general expression for resonance reactions⁴². If one considers the α_1 and p_1 fractions of the total fusion yield observed^{13,14} at $E_{\text{cm}} > 2.8$ MeV, the lower limits of the $\alpha_0 + \alpha_1$ and $p_0 + p_1$ contributions to the total cross-sections from the present experiment at the highest energies are 0.85 ± 0.07 and 0.68 ± 0.06 , respectively. However, the number of accessible excited states for both ^{20}Ne and ^{23}Na already reduces to half while moving from $E_{\text{cm}} = 2.8$ MeV to 1.5 MeV and the cross-sections for ^{20}Ne and ^{23}Na excited states drop more steeply than those for ground states, owing to the sharper decrease (by orders of magnitude) of the corresponding penetration factors. Monitoring the decrease of the penetration factors for the relevant states, and according to the results¹³ at $E_{\text{cm}} < 3$ MeV, the contribution to the total fusion yield from α and p channels other than $\alpha_{0,1}$ and $p_{0,1}$ was neglected in the modified R-matrix analysis within uncertainties at $E_{\text{cm}} < 2$ MeV lower than 1% and 2% for the α and p channels, respectively.

For all of the states involved in the procedure, the total widths are known and in several cases one of the partial widths (usually the α_0 partial width). Thus, the normalization constant NF in equation (3) and the missing partial widths were the only free parameters to match the modified R-matrix calculations with the indirect data for the four channels. Each calculated cross-section was folded with a Gaussian function having $\sigma = 30$ keV to account for energy resolution, as calculated from the beam spot size and divergence, the position-sensitive E-detector intrinsic energy and angle resolution, energy and angular straggling in target and dead layers. Total and partial widths and related uncertainties resulting from the fit are listed in Extended Data Table 2 for all levels entering the calculation. Uncertainties account for the error budget affecting experimental data (statistical and from background subtraction, when applicable) and correlation among the resonances in the four reaction channels and range from 10% to about 20%.

Astrophysical $S(E)$ factor. This factor is introduced to remove the dominant energy dependence of the cross-section between charged particles at astrophysical energies that is due to Coulomb barrier penetration. The $S(E)$ factor (in units of MeV b) is defined through the relationship:

$$S(E) = E\sigma(E)\exp(2\pi\eta) \quad (6)$$

where E is the incident energy in the centre-of-mass system, $\sigma(E)$ is the energy-dependent cross-section and $\exp(2\pi\eta)$ is the inverse of the Gamow factor, with η the Sommerfeld parameter, $\eta(E) = Z_1Z_2\alpha(\mu^2/2E)^{1/2}$ (where Z_1, Z_2 are the charges of the colliding nuclei, α is the fine structure constant, μ is the reduced mass in atomic mass units and c is the velocity of light).

For s-wave non-resonant reactions, the $S(E)$ factor is nearly independent of energy and it is the conventional quantity used to extrapolate to low energies.

For the $^{12}\text{C} + ^{12}\text{C}$ reaction, it is customary to use the so-called modified $S(E)$ factor, $S(E)^*$, which displays resonances more clearly. It is defined as:

$$S(E)^* = E\sigma(E) \times \exp(87.12E^{-1/2} + 0.46E) \quad (7)$$

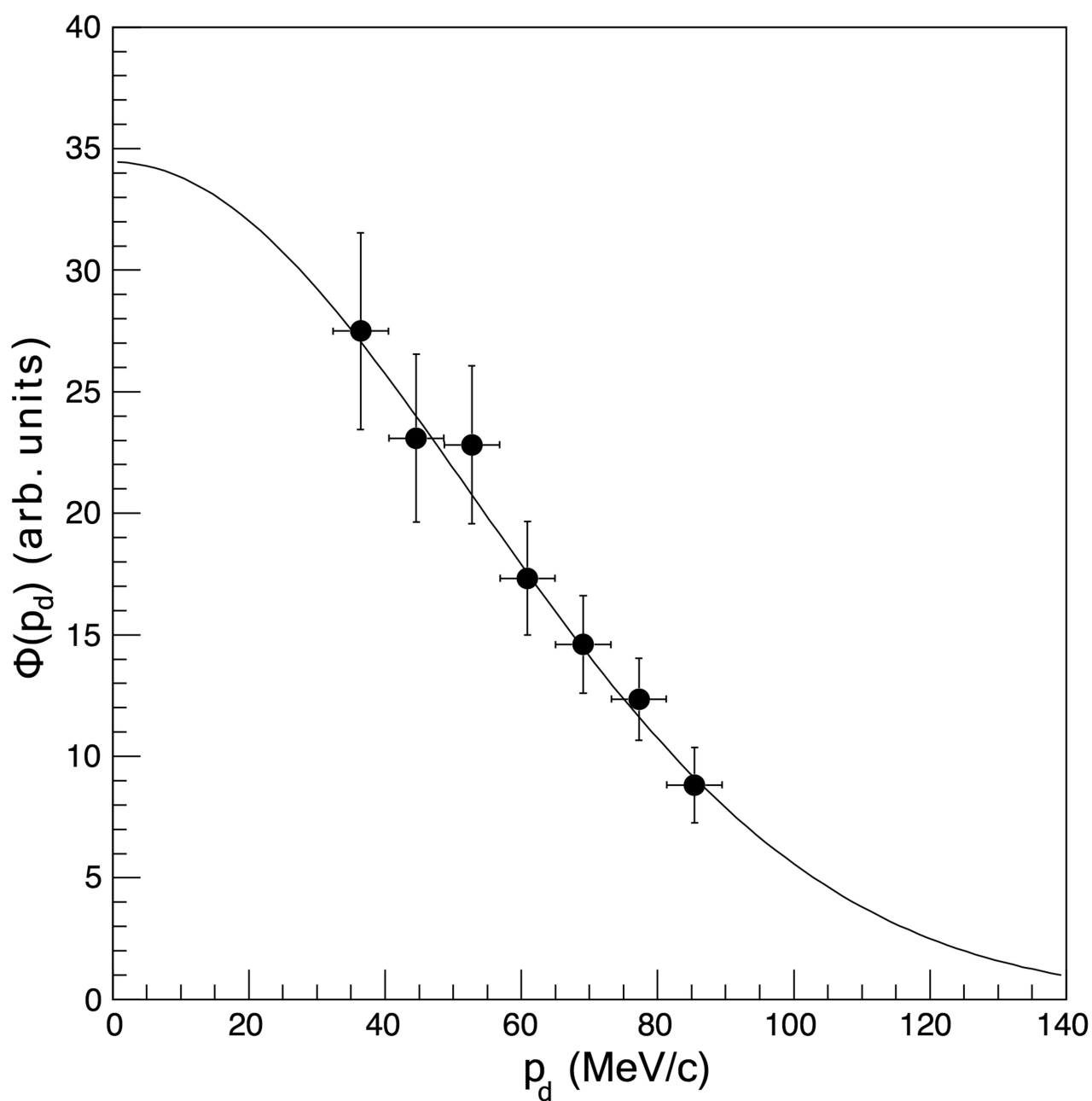
where the exponential term is the inverse of the Gamow factor with a correction arising from the second term in the Coulomb barrier approximation⁴⁵. In particular, the numerical factor 0.46 is the value of the size factor $g = 1/3(M_1M_2/(M_1 + M_2)R^3/2Z_1Z_2)^{1/2}$, with R_0 the nuclear separation and M_1, M_2 the masses of the colliding nuclei.

Numerical values of $^{12}\text{C} + ^{12}\text{C}$ reaction rate. Since the total $^{12}\text{C} + ^{12}\text{C}$ fusion yield at $E_{\text{cm}} < 2.8$ MeV is likely to be exhausted by the $\alpha_{0,1}$ and $p_{0,1}$ channels (see Methods section 'Numerical values of the $^{12}\text{C} + ^{12}\text{C}$ reaction rate'), we assume that the sum of their reaction rates in the E_{cm} range investigated here is representative of the total one. The numerical values of the total $^{12}\text{C} + ^{12}\text{C}$ reaction rate are given in Extended Data Table 2 expressed in units of $\text{cm}^3 \text{mol}^{-1} \text{s}^{-1}$ at temperatures of $T = 0.1 - 3$ GK. The lower and upper limits are computed using the total uncertainty derived by combining the rate uncertainties in quadrature for the four channels investigated. Each channel propagates the uncertainty in the THM $S(E)^*$ factor. The last column of Extended Data Table 2 shows the exponents of the power of ten factor multiplying the three previous columns.

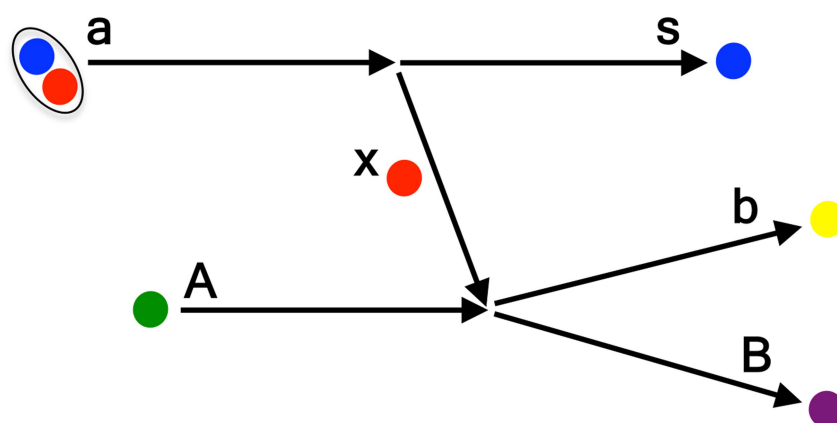
Data availability. All relevant data are available from the corresponding author on reasonable request. Data to calculate the rate ratio in Fig. 3 are included in Extended Data Table 2.

Code availability. We have not made publicly available the code for the modified R-matrix calculation because it is not intended for open use. However, it is available from the corresponding author upon request.

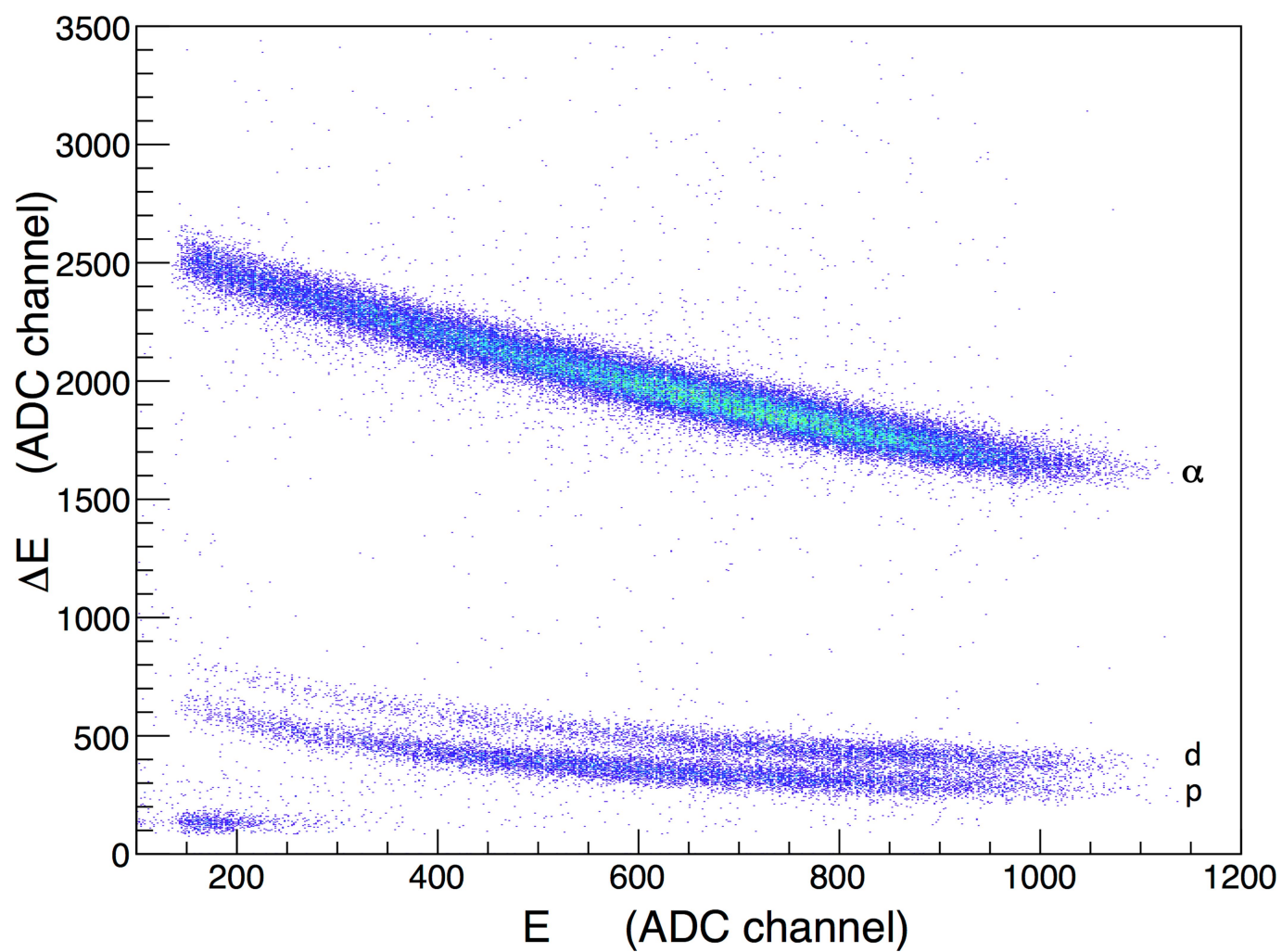
29. Strieder, F., Rolfs, C., Spitaleri, C. & Corvisiero, P. Electron-screening effects on fusion reactions. *Naturwissenschaften* **88**, 461–467 (2001).
30. Spitaleri, C. et al. Indirect $^7\text{Li}(p, \alpha)^4\text{He}$ reaction at astrophysical energies. *Phys. Rev. C* **60**, 055802–055809 (1999).
31. Lamia, L. et al. An updated $^6\text{Li}(p, \alpha)^3\text{He}$ reaction rate at astrophysical energies with the Trojan horse method. *Astrophys. J.* **768**, 65–73 (2013).
32. Tumino, A. et al. New determination of the $^2\text{H}(d, p)^3\text{H}$ and $^2\text{H}(d, n)^3\text{He}$ reaction rates at astrophysical energies. *Astrophys. J.* **785**, 96–113 (2014).
33. Pizzone, R. G. et al. First measurement of the $^{19}\text{F}(\alpha, p)^{22}\text{Ne}$ reaction at energies of astrophysical relevance. *Astrophys. J.* **836**, 57–63 (2017).
34. Spitaleri, C. et al. Measurement of the $^{10}\text{B}(p, \alpha)^7\text{Be}$ cross section from 5 keV to 1.5 MeV in a single experiment using the Trojan horse method. *Phys. Rev. C* **95**, 035801–035814 (2017).
35. Mukhamedzhanov, A. M. et al. Trojan Horse as an indirect technique in nuclear astrophysics. Resonance reactions. *J. Phys. G* **35**, 014016–014022 (2008).
36. La Cognata, M. et al. On the measurement of the $^{13}\text{C}(\alpha, n)^{16}\text{O}$ S-factor at negative energies and its influence on the s-process. *Astrophys. J.* **777**, 143–164 (2013).
37. La Cognata, M. et al. Effect of high-energy resonances on the $^{18}\text{O}(p, \alpha)^{15}\text{N}$ reaction rate at AGB and post-AGB relevant temperatures. *Astrophys. J.* **723**, 1512–1522 (2010).
38. Mukhamedzhanov, A. M. et al. Theory of deuteron stripping. From surface integrals to generalized R-matrix approach. *Phys. Rev. C* **84**, 044616–044622 (2011).
39. La Cognata, M. et al. The fluorine destruction in stars: first experimental study of the $^{19}\text{F}(p, \alpha)^{16}\text{O}$ reaction at astrophysical energies. *Astrophys. J.* **739**, L54–L60 (2011).
40. Zurmühle, R. W. et al. Observation of ^{12}C cluster transfer by angular correlation measurements. *Phys. Rev. C* **49**, 2549–2554 (1994).
41. Belyaeva, T. L., Zelenskaya, N. S. & Aguero Granados, M. Investigation of quasimolecular states in $^{24}\text{Mg}^*\text{Mg}$ through the analysis of the angular $d\alpha$ correlations in the $^{12}\text{C}(^{14}\text{N}, d)^{24}\text{Mg}(\alpha)^{20}\text{Ne}$ reaction. *Phys. At. Nucl.* **65**, 1616–1627 (2002).
42. Blatt, J. M. & Biedenharn, L. C. The angular distribution of scattering and reaction cross sections. *Rev. Mod. Phys.* **24**, 258–272 (1952).
43. Jiang, C. L. et al. Origin and consequences of $^{12}\text{C} + ^{12}\text{C}$ fusion resonances at deep sub-barrier energies. *Phys. Rev. Lett.* **110**, 072701–072705 (2013).
44. La Cognata, M. et al. A Trojan Horse approach to the production of ^{18}F in novae. *Astrophys. J.* **846**, 65–71 (2017).
45. Patterson, J. R., Winkler, H. & Zaidins, C. S. Experimental investigation of the stellar nuclear reaction $^{12}\text{C} + ^{12}\text{C}$ at low energies. *Astrophys. J.* **157**, 367–373 (1969).



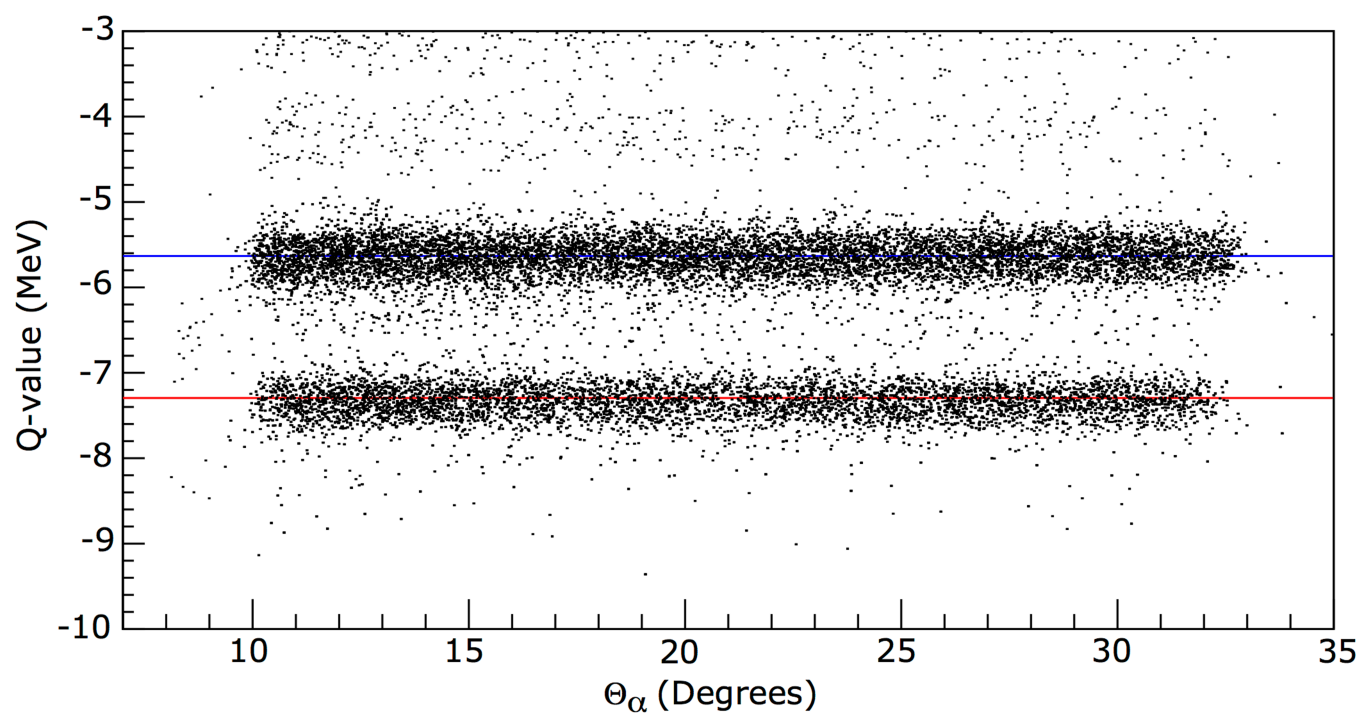
Extended Data Fig. 1 | Deuteron momentum distribution. The experimental distribution $\Phi(p_d)$ is shown as filled black circles. Error bars represent standard 1σ uncertainties. The black line represents the theoretical shape (see text for details).



Extended Data Fig. 2 | Pole diagram describing the quasi-free mechanism in the $A(a,bB)s$ reaction. The upper vertex refers to the break-up of a and the lower vertex shows the $A(x,b)B$ process. Colours help to highlight the role of individual particles in the mechanism.



Extended Data Fig. 3 | Typical ΔE - E spectrum. The strongest loci from the bottom to the top correspond to p , d and α . ADC, analogue-to-digital converter.



Extended Data Fig. 4 | Q-value as a function of the α detection angle Θ_α for the $^{12}\text{C}(^{14}\text{N}, \alpha ^{20}\text{Ne})^2\text{H}$ reaction. Blue and red solid lines cross the Q-value axis at -5.65 MeV and -7.28 MeV, highlighting the contributions of the ground and first excited states, respectively.

Extended Data Table 1 | Resonance parameters of ^{24}Mg levels entering the R -matrix fit and total plus partial widths resulting from the fit

$E_{\text{c.m.}}$ (MeV)	E_x (MeV)	J^π	$\Gamma_{\text{c.m.}}$ (keV)	$\Gamma_{\alpha 0}$ (keV)	$\Gamma_{\alpha 1}$ (keV)	Γ_{p0} (keV)	Γ_{p1} (keV)
2.664	16.597	4^+	29.3±6.1	10.7±2.2	0.90±0.18	1.1±0.2	16.6±3.5
2.567	16.500	3^-	42.8±4.8	4.8±0.9	0.8±0.1	0.7±0.1	36.5±3.7
2.537	16.470	6^+	6.91±0.95	2.0±0.4	0.31±0.05	0.7±0.1	3.9±0.4
2.5	16.433	7^-	9.6±1.1	0.9±0.2	0.019±0.003	0.07±0.01	8.6±0.9
2.455	16.388	2^+	37.5±5.4	15.1±2.7	0.31±0.04	4.1±0.8	18.0±1.9
2.403	16.336	4^+	22.4±2.8	4.6±0.8	0.7±0.1	1.7±0.3	15.4±1.6
2.369	16.302	8^+	0.49±0.07	0.18±0.03	0.032±0.005	0.06±0.01	0.22±0.02
2.338	16.271	4^+	37.9±5.6	13.6±2.5	3.5±0.5	4.6±0.9	16.2±1.7
2.263	16.196	6^+	7.3±1.2	3.3±0.6	0.45±0.07	1.3±0.3	2.2±0.2
2.196	16.129	3^-	31.6±4.3	1.0±0.2	3.0±0.5	8.8±1.7	18.8±1.9
2.095	16.03	2^+	7.2±1.1	1.1±0.3	1.00±0.02	3.4±0.4	1.7±0.4
2.038	15.971	3^-	43.9±4.4	15.2±2.0	0.10±0.02	2.1±0.5	26.5±1.8
1.946	15.879	4^+	43.4±7.6	4.0±0.7	3.6±0.5	27.4±5.5	8.4±0.9
1.888	15.821	3^-	87±16	5.3±0.9	8.1±1.3	65±13	8.0±0.9
1.853	15.786	4^+	12.8±1.9	2.4±0.4	1.3±0.2	3.8±0.8	5.2±0.5
1.777	15.71	4^+	30.6±4.6	9.0±1.6	0.45±0.07	8.0±1.6	13.1±1.3
1.747	15.68	0^+	4.8±0.6	0.38±0.07	0.10±0.02	1.0±0.2	3.3±0.3
1.671	15.604	2^+	29.0±4.2	6.3±1.1	1.9±0.3	7.7±1.5	13.1±1.3
1.593	15.526	6^+	23.7±3.3	8.3±1.5	0.40±0.06	1.3±0.3	13.7±1.4
1.544	15.477	2^+	17.1±2.1	1.0±0.2	0.07±0.01	3.5±0.7	12.5±1.2
1.503	15.436	2^+	12.6±1.9	1.7±0.3	0.7±0.1	4.6±0.9	5.7±0.6
1.445	15.378	4^+	31.9±4.7	10.4±1.9	2.2±0.3	5.2±1.1	14.0±1.4
1.414	15.347	4^+	16.6±2.5	8.0±1.4	0.10±0.01	2.3±0.5	6.2±0.6
1.37	15.31	5^-	9.1±1.5	0.07±0.01	0.02±0.003	5.5±1.1	3.5±0.4
1.293	15.226	4^+	25.1±4.2	0.6±0.1	0.9±0.1	16.0±3.2	7.6±0.8
1.274	15.207	5^-	39.3±6.4	26.9±4.9	0.6±0.08	2.7±0.5	9.2±0.9
1.239	15.172	4^+	60.8±9.3	28.1±5.1	7.0±1.1	5.1±1.0	20.6±2.1
1.201	15.134	4^+	15.7±1.8	0.49±0.09	0.30±0.04	2.0±0.4	12.9±1.3
1.175	15.108	4^+	16.6±2.5	1.6±0.3	2.8±0.4	5.6±1.1	6.7±0.7
1.055	14.988	5^-	10.4±1.8	0.5±0.1	0.8±0.3	6.0±1.1	3.1±0.3
0.988	14.921	0^+	12.9±1.8	2.1±0.5	0.40±0.03	2.5±0.6	7.9±0.7
0.909	14.842	2^+	14.9±1.8	2.1±0.4	1.4±0.2	7.8±0.9	3.6±0.3
0.877	14.81	1^-	5.6±0.6	2.1±0.2	0.10±0.02	3.0±0.3	0.40±0.08
0.805	14.738	4^+	8.6±1.3	2.2±0.3	3.6±0.4	2.7±0.6	0.11±0.02

Quoted uncertainties are $\pm 1\sigma$. Parameters are the centre-of-mass energy $E_{\text{c.m.}}$, the excitation energy E_x , the spin and parity J^π , the total width $\Gamma_{\text{c.m.}}$, the α_0 partial width $\Gamma_{\alpha 0}$, the α_1 partial width $\Gamma_{\alpha 1}$, the p_0 partial width Γ_{p0} and the p_1 partial width Γ_{p1} .

Extended Data Table 2 | Reaction rate of the $^{12}\text{C} + ^{12}\text{C}$ fusion reaction

Reaction Rate ($\text{cm}^3 \text{s}^{-1} \text{mol}^{-1}$)				
T (GK)	Adopted	Lower	Upper	Power
0.10	2.77	2.77	2.77	-52
0.11	7.80	7.69	7.91	-50
0.12	1.71	1.52	1.89	-47
0.13	4.59	3.55	5.64	-45
0.14	0.90	0.66	1.14	-42
0.15	0.97	0.71	1.24	-40
0.16	6.00	4.33	7.66	-39
0.18	5.85	4.21	7.49	-36
0.20	1.46	1.05	1.86	-33
0.25	3.12	2.25	3.99	-29
0.30	2.62	1.89	3.36	-26
0.35	3.91	2.82	5.01	-24
0.40	3.05	2.19	3.90	-22
0.45	1.95	1.40	2.49	-20
0.50	7.65	5.51	9.80	-19
0.60	2.49	1.79	3.19	-16
0.70	1.90	1.37	2.44	-14
0.80	5.97	4.30	7.63	-13
0.90	1.05	0.76	1.34	-11
1.00	1.24	0.91	1.56	-10
1.25	1.88	1.53	2.22	-08
1.50	1.03	0.93	1.11	-06
1.75	2.79	2.69	2.89	-05
2.00	4.41	4.35	4.47	-04
2.50	3.34	3.33	3.35	-02
3.00	0.86	0.86	0.86	00

The recommended value, 1σ lower and upper limits were computed at $T = 0.1\text{--}3$ GK covering the relevant astrophysical region. In the last column, the exponents of the power of ten multiplying columns 2, 3 and 4 are given.

Magnetic edge states and coherent manipulation of graphene nanoribbons

Michael Slota^{1,2}, Ashok Keerthi³, William K. Myers², Evgeny Tretyakov⁴, Martin Baumgarten³, Arzhang Ardavan^{2,5}, Hatef Sadeghi⁶, Colin J. Lambert⁶, Akimitsu Narita³, Klaus Müllen³ & Lapo Bogani^{1,2*}

Graphene, a single-layer network of carbon atoms, has outstanding electrical and mechanical properties¹. Graphene ribbons with nanometre-scale widths^{2,3} (nanoribbons) should exhibit half-metallicity⁴ and quantum confinement. Magnetic edges in graphene nanoribbons^{3,6} have been studied extensively from a theoretical standpoint because their coherent manipulation would be a milestone for spintronic⁷ and quantum computing devices⁸. However, experimental investigations have been hampered because nanoribbon edges cannot be produced with atomic precision and the graphene terminations that have been proposed are chemically unstable⁹. Here we address both of these problems, by using molecular graphene nanoribbons functionalized with stable spin-bearing radical groups. We observe the predicted delocalized magnetic edge states and test theoretical models of the spin dynamics and spin–environment interactions. Comparison with a non-graphitized reference material enables us to clearly identify the characteristic behaviour of the radical-functionalized graphene nanoribbons. We quantify the parameters of spin–orbit coupling, define the interaction patterns and determine the spin decoherence channels. Even without any optimization, the spin coherence time is in the range of microseconds at room temperature, and we perform quantum inversion operations between edge and radical spins. Our approach provides a way of testing the theory of magnetism in graphene nanoribbons experimentally. The coherence times that we observe open up encouraging prospects for the use of magnetic nanoribbons in quantum spintronic devices.

Theory predicts that graphene nanoribbons (GNRs) can have magnetic edges⁶ that display ferromagnetism and excellent spin-filtering properties⁷, in addition to interesting quantum-coherence features⁸. However, most GNRs do not have atomically precise edges, and bare graphene terminations are very sensitive to chemical modification⁹, so the properties of magnetic edge states—and even whether they exist—is still uncertain. Previous results based on microscopy have been largely blind to the magnetic effects. We have developed a bottom-up molecular synthesis, which allows for the fabrication of atomically precise GNRs with various structures that can be defined uniquely by the shape of molecular precursors^{10,11}. We have previously synthesized pure zigzag GNRs with localized edge states in ultrahigh vacuum, but magnetic characterization was challenging owing to chemical instability, so the spin properties of well-defined zigzag GNRs remain largely unexplored^{11,12}.

We overcome these challenges by injecting a spin density into the edge states of stable molecular GNRs synthesized via solution-based bottom-up chemical methods, using nitronyl nitroxide (NIT) radicals¹³ as magnetic injectors. The advantages of this approach are: that the groups introducing a magnetic functionality into the GNRs are well known¹³ and display interesting quantum properties¹⁴; that we do not rely on unobserved chemistry to create edge states; that the sample can be mass-produced, instead of appearing on only one device; and that we can test the classical and quantum spin properties in bulk samples.

The synthesis of NIT-radical-functionalized GNRs (NIT-GNRs) starts with Diels–Alder polymerization of a bromo-functionalized tetraphenylcyclopentadienone-based monomer (**1**), yielding a bromo-substituted precursor polymer (**2**; Fig. 1a). Palladium-catalysed cross-coupling of **2** to triphenylphosphine-gold(I)-(NIT-2-ide) yields magnetic NIT-polyphenylene, which provides a non-graphitized reference material (Fig. 1b). Graphitization of **2** yields the bromo-substituted nanoribbons (**3**), which are magnetically functionalized to NIT-GNRs by partial bromine substitution via cross-coupling (Fig. 1c)¹⁵. Size-exclusion chromatography of **2** yields an average molecular weight of 126 kg mol^{−1}, which corresponds to an average nanoribbon length of $\bar{l} > 100$ nm. Fourier-transform infrared, Raman and ultraviolet–visible spectroscopies corroborate the well-defined NIT-GNR structure, as in previous reports¹⁵, without an appreciable presence of transition-metal magnetic impurities (Supplementary Information).

The unpaired electron of the NIT resides in a π orbital that extends over two N–O groups and a C atom and overlaps considerably with the π orbitals of the aromatic backbone; the resulting spin density is injected efficiently into aromatic substituents (Supplementary Information)¹⁶. Modelling of NIT-GNRs using density functional theory reveals a sizeable spin density injected into the graphene backbone, which creates localized, non-dispersive states and a magnetic dispersive edge state, whereas the spins of the NIT-polyphenylene remain in completely localized states (Fig. 1b, c, Supplementary Information).

We can observe and manipulate the spin states directly using electron spin resonance¹⁷ (ESR), whereby the spin levels are split by a magnetic field and transitions are induced by microwave absorption. Static spectra at different frequencies (Fig. 2a) are reproduced using the spin Hamiltonian

$$H = H_Z + H_{\text{Hy}} + H_D + H_{\text{Ex}}$$

where the Zeeman term $H_Z = \mu_B \mathbf{B} \mathbf{g} \mathbf{S}_i$ contains the effect of the magnetic field \mathbf{B} on the i th spin \mathbf{S}_i via the Landé tensor \mathbf{g} (with μ_B the Bohr magneton); $H_{\text{Hy}} = \sum_{i,n} \mathbf{S}_i \mathbf{A}_{in} \mathbf{I}_n$ is the hyperfine interaction between the electron spin \mathbf{S}_i and the spin of the n th nucleus \mathbf{I}_n , mediated by the hyperfine interaction \mathbf{A}_{in} ; $H_D = \sum_{i,j} \mathbf{S}_i \mathbf{D}_{ij} \mathbf{S}_j$ is the dipolar coupling, with $\mathbf{D}_{ij} = \frac{g_i g_j \mu_B^2}{4\pi r_{ij}^3} \text{diag}(-1, -1, 2)$, which contains the vacuum permeability μ_0 and the spin–spin distance r_{ij} ; and $H_{\text{Ex}} = \sum_{i,j} \mathbf{S}_i \mathbf{J}_{ij} \mathbf{S}_j$ represents the exchange term produced by the exchange coupling \mathbf{J} . The parameters that best reproduce all frequencies are: $\mathbf{g} = \text{diag}(2.0097(5), 2.0060(4), 2.0026(1))$; hyperfine coupling with the ¹⁴N atoms $\mathbf{A}_N = \text{diag}(0.0, 3(2), 34(2))$ MHz, tilted by $\varphi = 9^\circ$ in-plane relative to \mathbf{g} ; and $D_{12} \equiv D_1 = 11.0 \pm 0.5$ MHz and $D_{13} = D_{23} \equiv D_2 = 8.5 \pm 0.5$ MHz for the along- and across-edge interactions, respectively (Fig. 2b). Within error, the same results are obtained for the radicals on NIT-GNRs: $\mathbf{g} = \text{diag}(2.0098(5), 2.0059(5), 2.0026(1))$, $\mathbf{A}_N = \text{diag}(0.0, 5(2), 34(2))$, $D_1 = 11.0 \pm 0.5$ MHz and $D_2 = 8.5 \pm 0.5$ MHz. The inter- and intra-

¹Department of Materials, University of Oxford, Oxford, UK. ²Centre for Advanced ESR, University of Oxford, Oxford, UK. ³Max-Planck-Institut für Polymerforschung, Mainz, Germany. ⁴N. N. Vorozhtsov Novosibirsk Institute of Organic Chemistry, Novosibirsk, Russia. ⁵Clarendon Laboratory, University of Oxford, Oxford, UK. ⁶Quantum Technology Centre, Physics Department, Lancaster University, Lancaster, UK. *e-mail: lapo.bogani@materials.ox.ac.uk

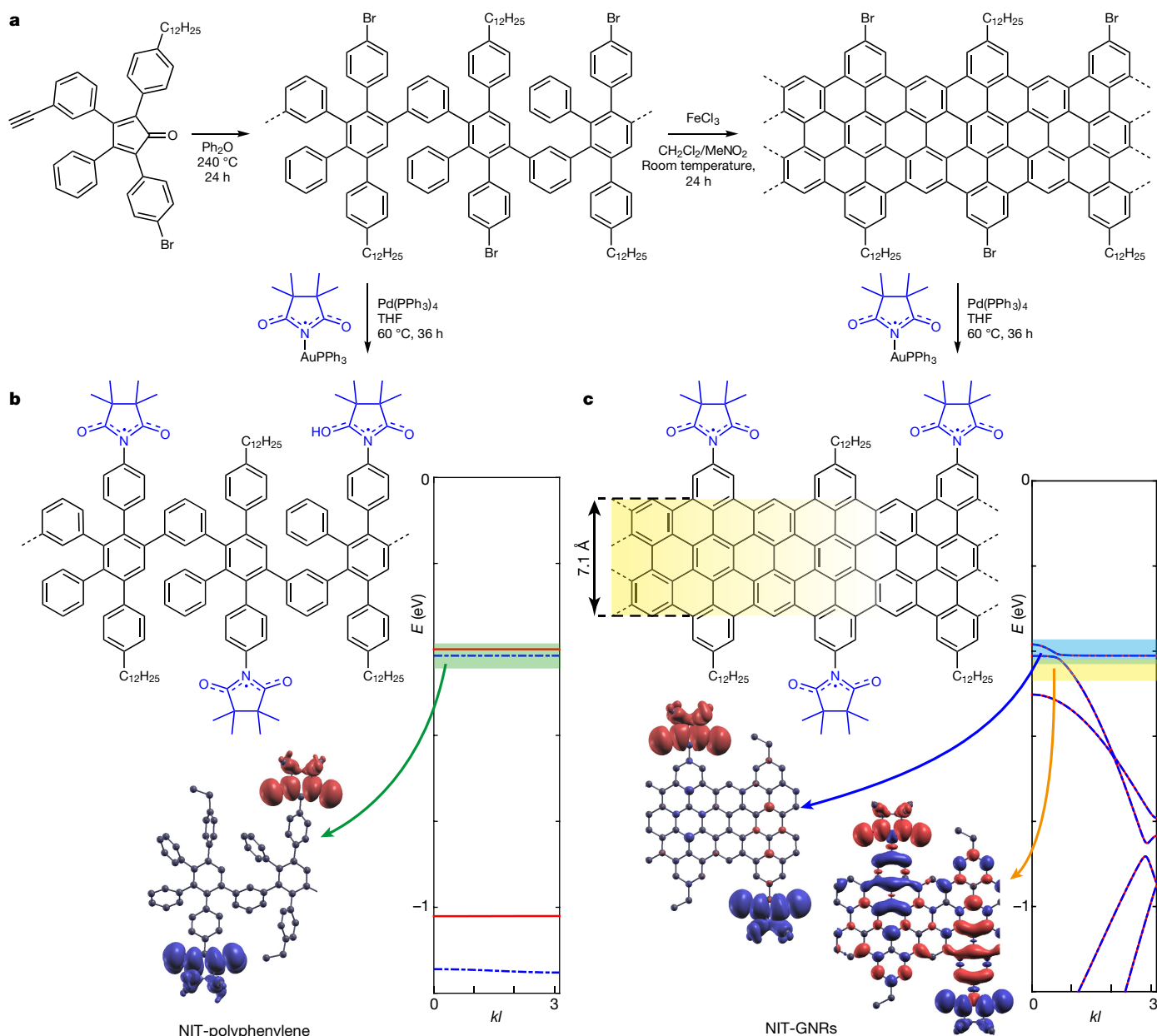


Fig. 1 | Functionalized graphene nanoribbons. **a**, Synthetic path to NIT-polyphenylene and NIT-GNRs; spin-bearing radicals are shown in blue. **b**, NIT-polyphenylene and its local density of states E . The energy levels (as functions of the wavevector k times the repeating-unit length l) show no band structure and spins localized on the NIT groups. **c**, A NIT-GNR

edge exchange interactions are $J_1 = -25 \pm 5$ MHz and $J_2 = 12 \pm 3$ MHz, in agreement with the sign expected from theoretical predictions¹⁸ and Goodenough–Kanamori rules¹³. These signals are attributed to the spin density that is localized on the NIT radicals.

In addition to these signals, NIT-GNRs display the predicted edge state as a strong feature with uniaxial anisotropy: $g_{||} = 2.0024(3)$, $g_{\perp} = 2.0041(2)$. Metallic impurities would produce ESR line widths of tens of millitesla, rather than the 1–2 mT observed. Metals and spin-bearing defects in the graphene backbone would have a different hyperfine coupling from the NIT radicals and would not display all of the characteristics of NIT radicals, and double electron–electron resonance (DEER) experiments would not be possible with randomly placed impurities (see below).

The shape and line width of the ESR signal rule out magnetic impurities and are consistent with previous indications of delocalized spin states¹⁹, providing conclusive evidence for the existence of the edge spin states that have long been predicted for graphene nanoribbons^{4–7}.

and its band structure, showing localized states and spin injection inside delocalized edge states. In **b** and **c**, densities calculated for different energy ranges are depicted as green, blue and orange shaded areas (see also the arrows); the blue dashed and red solid lines correspond to local densities of spin-up and spin-down states, for a given energy interval.

Theory predicts that the honeycomb lattice of graphene introduces an axial spin–orbit effect, Δ_{SO} , whereas the breaking of the mirror symmetry of the plane produces a Rashba-type transverse term, Δ_R , yielding the Hamiltonian

$$H_{SO} = \pm \Delta_{SO} \sigma_z S_z + \Delta_R (\pm \sigma_x S_x - \sigma_y S_y)$$

where \pm denote the valley degrees of freedom, and S_i and σ_i are the spin and pseudospin Pauli matrices⁸, respectively. $\Delta_{SO} \approx 15 \mu\text{eV}$ and $\Delta_R \approx 1 \mu\text{eV}$ are extracted by considering that $|\Delta E(g_i - g_e)| = 2\Delta_p$, where g_e is the free-electron value, and perturbation theory is used to account for the effect of excited states at energy ΔE (available from the ab initio calculations). This constitutes direct experimental confirmation of tight-binding estimates of spin–orbit coupling in graphene^{20,21} and of its suppression compared to carbon nanotubes, as predicted by the lattice symmetry and the absence of curvature^{5,21}. These observations, together with the fact that the static spectra are largely insensitive to

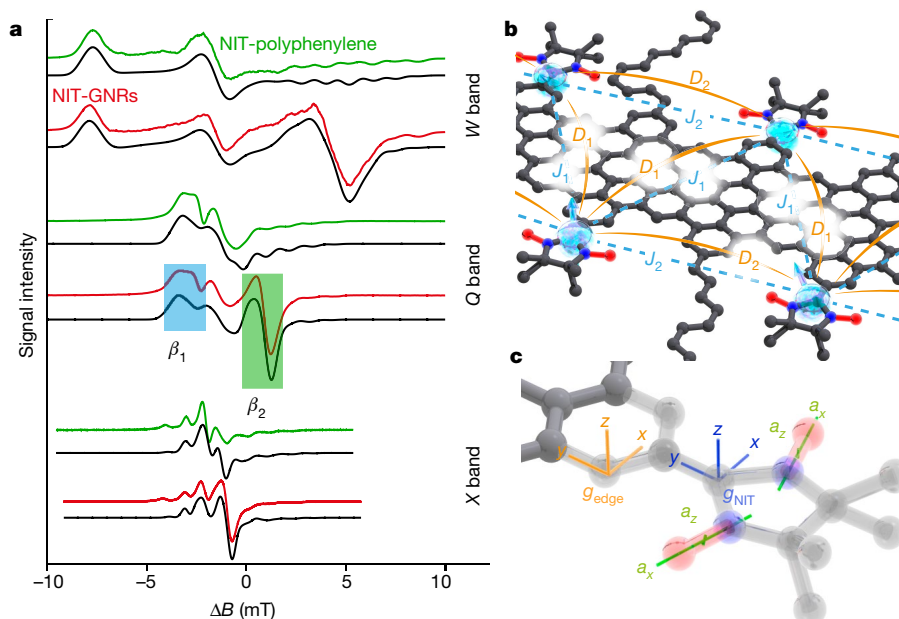


Fig. 2 | Static spectra and magnetic interaction pathways. **a**, Multi-frequency ESR spectra for NIT-polyphenylene (green) and NIT-GNRs (red), along with simulations (black), plotted against the magnetic field from the edge-state resonance. DEER probe and pump windows are labelled as β_1 and β_2 respectively. **b**, Interaction pathways for radical spins, showing exchange (J_1 and J_2 ; blue dashed lines) and dipolar (D_1 and D_2 ;

orange lines) interactions. **c**, Orientation of the ^{14}N hyperfine channel (green), with the lengths of the axes (a_x , a_y and a_z) proportional to the principal tensor elements (a_y is smaller than the width of the axis), and orientations of the local g -tensor frame of the radical (blue, g_{NIT}) and that of the graphene edge state (orange, g_{edge}).

exchange interactions, indicate that the NIT-GNRs fall into a very interesting regime, in which coherent manipulation of the spins is possible.

We therefore explore the quantum spin coherence using time-resolved ESR. The quantum evolution of a spin can be represented as a movement over the Bloch sphere, with zenith positions corresponding to pure $|1/2\rangle$ and $|-1/2\rangle$ states and all of their possible combinations mapped on the sphere (Fig. 3a). The spin relaxation time T_1 represents spin flips (vertical displacement), whereas the phase memory time T_2 describes the evolution of the quantum phase information (azimuthal movement). We measure T_1 using the ‘picket fence’ technique²² and T_m (a measure of the dephasing time) using the Hahn-echo decay. We fit the decay of the echo signal Y over time τ with an exponential function,

$$Y(\tau) = Y_0 e^{-(2\tau/T_m)^x} [1 + k_1 \sin(2\omega\tau + \varphi_1) + k_2 \sin(4\omega + \varphi_2)]$$

which includes a signal at zero time Y_0 , modulation by the environment at a frequency ω (amplitudes k_1 and k_2 for first- and second-order effects), the phases φ_1 and φ_2 , and a stretch parameter x (Fig. 3b). We always find $x = 1$, which indicates that the approximation of the relaxation time is good, that successive events are uncorrelated and that $T_m \approx T_2$, as traditionally defined²³.

The T_1 values (roughly 10^{-5} s) validate theoretical predictions²⁴. The temperature dependence of $1/T_1$ (Fig. 3c) shows three main regimes: a linear one below 25 K, which is characteristic of spin–phonon energy transfer; a Raman region between 25 K and 200 K, in which relaxation happens via virtual states; and a room-temperature region in which local vibrational modes have a role, with the same characteristic energy ($1,354\text{ cm}^{-1}$) for NIT-GNR and NIT-polyphenylene, which we tentatively assign to the N–O stretching mode. Theories of low-temperature spin–phonon relaxation in graphene^{5,20} quantum dots consider a deformation-potential mechanism, which is active for longitudinal acoustic phonons, and a bond-length-change mechanism, which is active for transversal and longitudinal acoustic modes. These mechanisms, in conjunction with the absence of Van Vleck cancellation²⁰, are predicted to generate the linear dependence that we observe here at low fields. The other hypothesized mechanism—spin-state admixture⁵—can be ruled out by the temperature and field dependences that we observe

and by the low value of the Rashba spin–orbit coupling, to which it is linked by symmetry selection rules^{5,20}.

Even without any optimization, NIT-GNRs display $T_m = 0.5\text{ }\mu\text{s}$ at room temperature and $T_m = 1.1\text{ }\mu\text{s}$ at 85 K (Fig. 3c, Supplementary Information)—100 times longer than the 12 ns available in spintronic devices²⁵. These high values are probably linked to the efficient suppression of scattering in atomically regular edges. NIT-GNRs exhibit only a slight increase in T_m at lower temperatures, whereas NIT-polyphenylene exhibits a minimum at 170 K and a broad maximum at 60 K, attributable to the progressive freezing of the benzene–benzene σ bonds in the backbone. Although T_m for the localized radicals in NIT-polyphenylene might be slightly longer, the NIT-GNRs enable us to validate theories of spin relaxation in graphene, have an edge state that is connected to transport and are promising for quantum operations.

We now determine the sources of decoherence in NIT-GNRs. The modulation of the Hahn-echo amplitude (Fig. 3b) at $\omega/(2\pi) = 3.6\text{ MHz}$ —a frequency typical of ^{13}C spin–nuclei interactions—suggests that hyperfine decoherence channels are important. Electron–electron double-resonance-detected nuclear magnetic resonance (EDNMR) allows us to deconvolute the different nuclear contributions^{16,23} (Fig. 4a). ^{14}N coupling is dominant, which confirms the analysis of continuous-wave spectra, but ^{13}C , ^1H single quantum transitions, ^{14}N , ^{13}C double quantum transitions and nitrogen–carbon mixed transitions also have important roles. The coupling strength to the ^{13}C of the graphene backbone (about 10 MHz) is considerably smaller than theoretical estimates for confined graphene dots^{5,26}, in which anisotropic, Fermi-contact and nucleus–orbital interactions contribute to a total ^{13}C hyperfine interaction of about 70 MHz. These couplings suggests that nuclei could be used as computational resources²⁷.

Finally, we consider the coupling between localized spins and the edge state. Information about electron–electron interactions is obtained by four-pulse DEER (Fig. 4b)^{17,22}, whereby the system is initialized and probed at the g_x resonance of the radicals and perturbed at the resonance condition of the edge state. The resulting spectrum displays an intriguing slow oscillation that is overlaid by fast ones (Fig. 4c). The fast period corresponds to the D_1 and D_2 interactions, which are too strong for accurate resolution using DEER and are better appreciated via the continuous-wave spectra. Slow oscillations correspond to interactions

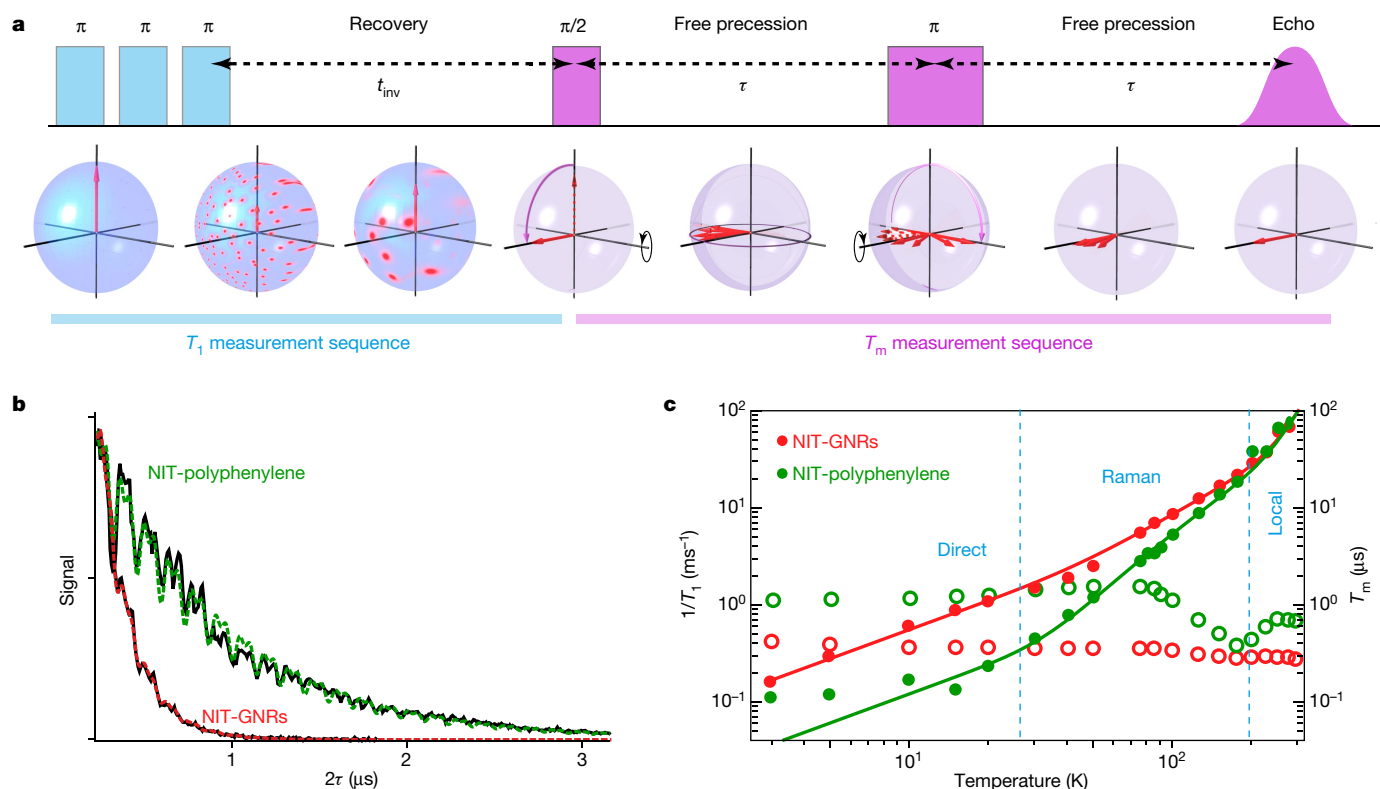


Fig. 3 | Spin-lattice relaxation and spin coherence times. **a**, Pulse sequence used to extract the spin relaxation times and the Bloch-sphere representation. A series of π pulses (blue) erases the spin polarization (red arrow), which recovers after a time T_1 during the free-evolution interval t_{inv} . The spins are then rotated to the x - y plane with a $\pi/2$ pulse (violet) and allowed to precess for a time τ . A π rotation in the middle of the free precession causes an echo signal when the spins regroup (violet bell

curve). **b**, X-band Hahn-echo signal versus the delay time τ for NIT-GNRs and NIT-polyphenylene, at 85 K (black lines). The red and green dashed lines are fits to the data, which yield T_m . **c**, $1/T_1$ (filled circles, left axis) and T_m (open circles, right axis) versus temperature for NIT-GNRs (red) and NIT-polyphenylene (green), at 9.4 GHz. The solid lines are simulations for $1/T_1$ (see text); the vertical blue dashed lines separate the different regimes of spin-lattice relaxation.

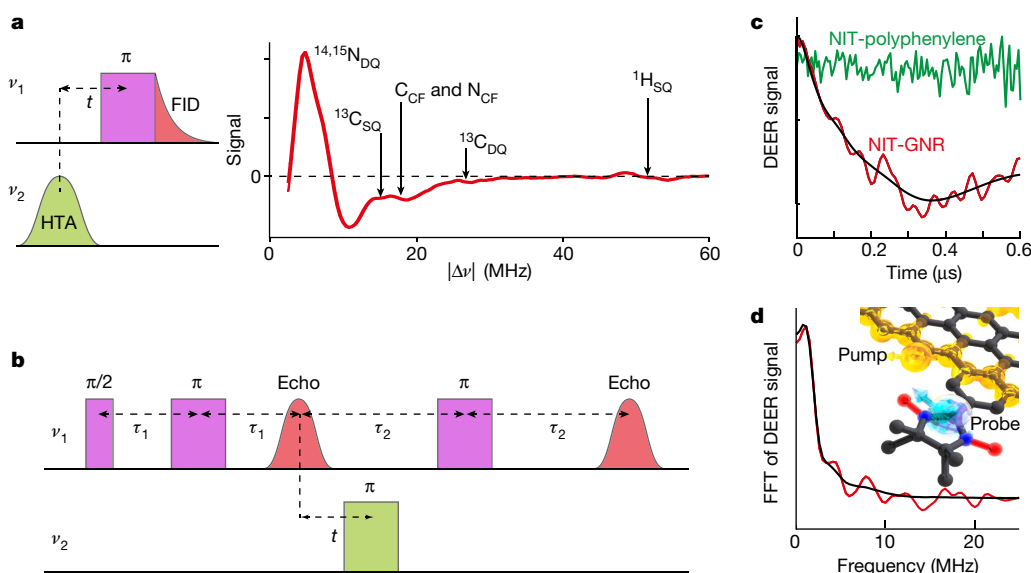


Fig. 4 | Hyperfine coupling and multi-spin operability in GNRs. **a**, EDNMR sequence, using a high-turning-angle pulse (HTA) at frequency ν_2 (the edge-state resonance β_2 ; Fig. 2a) while frequency ν_1 is swept and the free-induction decay (FID) is measured. The spectrum for the NIT-GNRs (Q band, 85 K) shows single quantum (SQ), double quantum (DQ) and combination frequency (CF) transitions. **b**, DEER sequence, with ν_1 set at the localized spin resonance and ν_2 at the edge state (β_1 and β_2 in Fig. 2a), used to determine the spin-spin interactions

and to perform quantum inversion operations between the edge and localized spins. **c**, Background-corrected time-domain DEER spectrum for NIT-polyphenylene (green) and NIT-GNRs (red). The black line singles out the low-frequency interactions. **d**, Fast Fourier transform (FFT) of the DEER signal of NIT-GNRs, showing the interaction energy spectrum that is characteristic of two-spin operations. The black line singles out the contribution from edges that interact with localized spins (see inset).

between localized and edge-state spins yielding a radical–edge spin interaction of 1.5 MHz (Fig. 4d); these oscillations are absent in NIT-polyphenylene, in agreement with the lack of edge states. The edge–radical spin inversion time that we extract (about 330 ns) is considerably shorter than T_m , enabling coherent inversion operations using graphene edge states and localized spins.

This finding, in conjunction with recent results on transport on molecular nanoribbons, could lead to fascinating possibilities: quantum operations can in principle be performed via single-electron transport, and the spin states detected electrically, potentially making our NIT-GNRs ideal candidates for quantum nanoelectronic devices. The interaction of multiple radical spins with a coherent, delocalized edge state could allow a single flowing electron to transmit entanglement along the spin ensemble⁸. Furthermore, such molecular nanoribbons are useful for testing fundamental theories of graphene. Our measurements of spin–orbit, hyperfine and edge–spin coupling already reveal physics that would otherwise be accessible only by overcoming the present challenges in studies of the quantum Hall effect at sub-millikelvin temperatures^{20,21}. Detailed access to the spin dynamics, together with an atomically defined structure, opens up a path to the quantitative analysis of spin–phonon interactions in graphene dots. The study of different molecular spin injectors and of different aromatic backbones¹⁰ (which can be used, for example, to modulate the spin coupling) provides the foundations for an area of chemistry that mixes molecular magnetism and graphene. Environmental effects, such as GNR–GNR or GNR–substrate interactions, are an interesting area of future research; calculations show, encouragingly, no detrimental effect on the spin density by deposition on hexagonal boron nitride (Supplementary Information). Because ¹⁴N hyperfine coupling is a dominant decoherence channel, there is ample room to increase T_2 , for example, by dynamic nuclear spin polarization²⁸, isotopic substitution²⁹ or chemical engineering³⁰. Full investigations of magnetic doping effects and of incomplete edge functionalization with radicals is currently underway. We expect the findings to reveal GNRs as powerful tools for investigating finite-size effects in quantum Heisenberg spin chains¹⁸.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0154-7>.

Received: 28 April 2017; Accepted: 14 February 2018;

Published online 30 May 2018.

1. Castro Neto, A. H., Guinea, F., Peres, N. M. R., Novoselov, K. S. & Geim, A. K. The electronic properties of graphene. *Rev. Mod. Phys.* **81**, 109–162 (2009).
2. Jiao, L., Zhang, L., Wang, X., Diankov, G. & Dai, H. Narrow graphene nanoribbons from carbon nanotubes. *Nature* **458**, 877–880 (2009).
3. Jia, X. et al. Controlled formation of sharp zigzag and armchair edges in graphitic nanoribbons. *Science* **323**, 1701–1705 (2009).
4. Son, Y. W., Cohen, M. L. & Louie, S. G. Half-metallic graphene nanoribbons. *Nature* **444**, 347–349 (2006). corrigendum 446, 342 (2007).
5. Recher, P. & Trauzettel, B. Quantum dots and spin qubits in graphene. *Nanotechnology* **21**, 302001 (2010).
6. Meunier, V., Souza Filho, A. G., Barros, E. B. & Dresselhaus, M. S. Physical properties of low-dimensional sp^2 -based carbon nanostructures. *Rev. Mod. Phys.* **88**, 025005 (2016).
7. Pesin, D. & MacDonald, A. H. Spintronics and pseudospintronics in graphene and topological insulators. *Nat. Mater.* **11**, 409–416 (2012).
8. Trauzettel, B., Bulaev, D. V., Loss, D. & Burkard, G. Spin qubits in graphene quantum dots. *Nat. Phys.* **3**, 192–196 (2007).
9. Barone, V., Hod, O. & Scuseria, G. E. Electronic structure and stability of semiconducting graphene nanoribbons. *Nano Lett.* **6**, 2748–2754 (2006).

10. Narita, A., Wang, X.-Y., Feng, X. & Müllen, K. New advances in nanographene chemistry. *Chem. Soc. Rev.* **44**, 6616–6643 (2015).
11. Ruffieux, P. et al. On-surface synthesis of graphene nanoribbons with zigzag edge topology. *Nature* **531**, 489–492 (2016).
12. Cai, J. et al. Atomically precise bottom-up fabrication of graphene nanoribbons. *Nature* **466**, 470–473 (2010).
13. Caneschi, A., Gatteschi, D. & Rey, P. The chemistry and magnetic properties of metal nitronyl nitroxide complexes. *Prog. Inorg. Chem.* **39**, 331–429 (1991).
14. Collauto, A. et al. A slow relaxing species for molecular spin devices: EPR characterization of static and dynamic magnetic properties of a nitronyl nitroxide radical. *J. Mater. Chem.* **22**, 22272–22281 (2012).
15. Narita, A. et al. Synthesis of structurally well-defined and liquid-phase-processable graphene nanoribbons. *Nat. Chem.* **6**, 126–132 (2014).
16. Zheludev, A. et al. Spin-density in a nitronyl nitroxide free-radical-polarized neutron-diffraction investigation and ab initio calculations. *J. Am. Chem. Soc.* **116**, 2019–2027 (1994).
17. Schweiger, A. & Jeschke, G. *Principles of Pulse Electron Paramagnetic Resonance* (Oxford Univ. Press, Oxford, 2001).
18. Golor, M., Wessel, S. & Schmidt, M. J. Quantum nature of edge magnetism in graphene. *Phys. Rev. Lett.* **112**, 046601 (2014).
19. Rao, S. S. et al. Spin dynamics and relaxation in graphene nanoribbons: electron spin resonance probing. *ACS Nano* **6**, 7615–7623 (2012).
20. Min, H. et al. Intrinsic and Rashba spin-orbit interactions in graphene sheets. *Phys. Rev. B* **74**, 165310 (2006).
21. Kane, C. L. & Mele, E. J. Z_2 topological order and the quantum spin Hall effect. *Phys. Rev. Lett.* **95**, 146802 (2005).
22. Eaton, G. R. & Eaton, S. S. in *Multifrequency Electron Paramagnetic Resonance: Theory and Applications* (ed. Misra, S. K.) Ch. 17 (Wiley, Weinheim, 2011).
23. Klauder, J. R. & Anderson, P. W. Spectral diffusion decay in spin resonance experiments. *Phys. Rev.* **125**, 912–932 (1962).
24. Struck, P. R. & Burkard, G. Effective time-reversal symmetry breaking in the spin relaxation in a graphene quantum dot. *Phys. Rev. B* **82**, 125401 (2010).
25. Drögeler, M. et al. Spin lifetimes exceeding 12 ns in graphene nonlocal spin valve devices. *Nano Lett.* **16**, 3533–3539 (2016).
26. Fischer, J., Trauzettel, B. & Loss, D. Hyperfine interaction and electron-spin decoherence in graphene and carbon nanotube quantum dots. *Phys. Rev. B* **80**, 155401 (2009).
27. Dutt, M. G. et al. Quantum register based on individual electronic and nuclear spin qubits in diamond. *Science* **316**, 1312–1316 (2007).
28. Foletti, S., Bluhm, H., Mahalu, D., Umansky, V. & Yacoby, A. Universal quantum control of two-electron spin quantum bits using dynamic nuclear polarization. *Nat. Phys.* **5**, 903–908 (2009).
29. Balasubramanian, G. et al. Ultralong spin coherence time in isotopically engineered diamond. *Nat. Mater.* **8**, 383–387 (2009).
30. Shiddiq, M. et al. Enhancing coherence in molecular spin qubits via atomic clock transitions. *Nature* **531**, 348–351 (2016).

Acknowledgements We thank the European Research Council (ERC-StG 338258 OptoQMol), the EU (COST-CA15128, MOLESCO-606728 and Graphene Flagship), the EPSRC (QuEEN grant), the Royal Society (University Research Fellowship and URF grant), the RFBR (17-53-50043), the Max Planck Society and the German DAAD Bilateral Exchange of Academics (2015/50015739) for financial support.

Reviewer information Nature thanks E. Coronado, D. Gatteschi, A.-P. Li and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.S. and W.K.M. performed the ESR characterization. A.K. and E.T. performed the synthesis and related characterization, for which M.B., A.N. and K.M. provided supervision. H.S. and C.J.L. performed the numerical modelling. M.S., W.K.M., A.A. and L.B. contributed to the ESR data analysis. M.S., A.N., K.M. and L.B. conceived the experiments and M.S. and L.B. wrote the manuscript. All authors contributed to the discussion and to the final version of the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0154-7>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to L.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Precursors. All chemicals were used as purchased without further purification. All reactions dealing with air- or moisture-sensitive compounds were carried out in a dry reaction vessel under argon. The starting materials S1, Poly-Br and GNR-Br were synthesized by adapting published protocols^{15,31,32}; the detailed synthesis is reported in Supplementary Information. Unless otherwise noted, materials were purchased from commercial suppliers (Fluka, Aldrich, Acros, ABCR, Merck and others) and used as received without further purification.

Material characterization. Analytical size-exclusion chromatography (SEC) was performed on SDV PSS GPC columns using THF as the eluent at a temperature of 303 K. Absorbance was determined on a UV S-3702 detector (SOMA) at a fixed wavelength of 270 nm. The samples were referenced with respect to standard polystyrene (PS) as well as poly(p-phenylene) (PPP) calibration curves. Solution UV-Vis absorption spectra were recorded at room temperature on a Perkin-Elmer Lambda 900 spectrophotometer. GNR samples were dispersed in N-methyl-pyrrolidone (NMP) by using sonication (30 min) in a Branson-1510 ultrasonicator followed by filtration through 5- μ m polytetrafluoroethylene (PTFE) filters. Infrared spectroscopy was measured on a Nicolet 730 FT-IR spectrometer equipped with an attenuated total reflection (ATR) set-up. The samples were deposited as pristine material on the diamond crystal and pressed on it with a stamp. Measurements with a scan number of 128 were recorded for each sample and the background was subtracted.

ESR spectroscopy. X-band (about 9.5 GHz) and W-band (about 94 GHz) ESR spectra were acquired using a Bruker EleXsys E680. Q-band ESR spectra were obtained on a Bruker EleXsys E580. X-band continuous-wave (CW) spectra were additionally recorded on a Bruker EMX. The sample temperature was maintained with an Oxford Instruments CF9350 cryostat and controlled with an Oxford Instruments ITC503. The microwave resonators that we used were Bruker ER4118X-MS-3W1 for X-band (Bruker E680), Bruker EN510702 for Q-band and Bruker EN600-1021H for W-band measurements. The sample was prepared as loose powder in low-background quartz glass tubes. X-band CW measurements were calibrated using polystyrene. High-field measurements at Q-band and W-band frequencies were calibrated using a MgO standard.

CW ESR. CW measurements were performed at X-band and Q-band frequencies at room temperature. The modulation amplitude was set to 1 G and the modulation frequency to 100 kHz for X-band and to 50 kHz for Q-band measurements. Spectra were simulated using the Matlab package *EasySpin*³³. Errors on the g , A and D values are obtained as standard deviations from multiple sets of measurements. Errors on J values are estimated from goodness of simulation.

FID-detected ESR. Free-induction-decay (FID)-detected ESR spectra at W-band frequencies were obtained by exciting the spin system with a long π pulse. This enables detectable FID signals to persist after the receiver protection switch. By integrating the signal against the field, the result comes very close to the integral of the corresponding CW ESR experiment³⁴. The measurement was performed at 85 K to ensure a strong signal. Spectra were simulated using the Matlab package *EasySpin*³³.

Determination of T_1 . In the picket-fence pulse sequence (Fig. 3b), which we used to determine the spin-lattice relaxation time T_1 , we set the $\pi/2$ and π pulse durations to 16 ns and 32 ns, respectively, and used a four-step phase-cycling procedure. The resulting data were fitted with a sum of two exponential functions, with one accounting for spin-lattice relaxation and the other for spectral diffusion processes. Spectral diffusion arises from electron-electron and electron-nuclear couplings, thus adding another relaxation mechanism. The picket-fence pulse sequence minimizes this contribution.

Fit of T_1 . We describe the spin-lattice relaxation by using three main mechanisms, which scale differently with temperature T . By considering direct relaxation, Raman processes and the excitation of local vibrational modes, we arrive at²²:

$$\frac{1}{T_1} = A_{\text{lin}}T + A_{\text{Ram}}\left(\frac{T}{\theta_D}\right)^9 \int_0^{\theta_D/T} \frac{x^8 e^x}{(e^x - 1)^2} dx + A_{\text{loc}} \left[\frac{e^{\Delta_{\text{loc}}/T}}{(e^{\Delta_{\text{loc}}/T} - 1)^2} \right]$$

where A_{lin} , A_{Ram} , A_{loc} denote the rate constants of the three processes, θ_D is the Debye temperature, x is an integration variable and Δ_{loc} is the excitation energy of a local vibrational mode. For both compounds, we find $\theta_D = 200$ K, $A_{\text{loc}} = 3.5 \times 10^7$ s⁻¹ and $\Delta_{\text{loc}} = 1,950$ K (1,354 cm⁻¹), in excellent agreement with the local excitation of the stretching modes of the N-O bond, as obtained by infrared

spectroscopy³⁵. Furthermore, we find $A_{\text{lin}} = 12$ s⁻¹, $A_{\text{Ram}} = 3.0 \times 10^{-16}$ s⁻¹ for NIT-polyphenylene and $A_{\text{lin}} = 55$ s⁻¹, $A_{\text{Ram}} = 2.2 \times 10^{-16}$ s⁻¹ for NIT-GNRs. The relaxation is affected by electron-electron interaction effects because our samples are highly spin-concentrated. This results in a linear region even above 10 K, which is not related to direct relaxation²².

Hahn echo. The Hahn-echo sequence (Fig. 3b) enables the determination of the phase memory time T_m , which represents an upper limit for T_2 , because several effects, such as instantaneous and spectral diffusion, affect the measurement. We set the pulse durations to 16 ns for the $\pi/2$ pulse and 32 ns for the π pulse, phased the signal and integrated over the second half of the echo. We used a 16-step phase-cycling procedure. We identified T_m to be equal to T_2 .

EDNMR spectroscopy. EDNMR was performed at Q-band frequencies and 85 K. Pulse shapes were formed using an arbitrary-waveform generator. The duration of the Gaussian-shaped excitation pulse was set to 1,200 ns, which allows for the excitation of forbidden nuclear transitions³⁶. The detection π pulse after a delay of t was set to 400 ns and the FID recorded. The central line around $\Delta\nu = 0$ was removed by using Voigtian and additional Lorentzian line shapes.

DEER spectroscopy. DEER measurements were performed at Q-band frequencies and 85 K. Pulses were created using an arbitrary-waveform generator. The durations of the detection pulses were set to 12 ns and 18 ns for NIT-GNRs and NIT-polyphenylene, respectively. The duration of the pump pulse (ELDOR pulse) was set to 20 ns. We integrated over the resulting echo against the variable time delay of the ELDOR pulse. The time-domain DEER data shown were obtained after normalization, long-pass filtering (23.5-MHz threshold) and subtraction of a three-dimensional background using the Matlab package *DeerAnalysis*³⁷.

Numerical calculations. Theoretical modelling was performed using the Gaussian³⁸ and SIESTA³⁹ implementations of density functional theory. For the spin density calculations of the two systems shown in Fig. 1, Gaussian was used with a *ulsd/a-6-311++g(d,p)* functional and basis set and the XQC method for the self-consistent reaction field. We found a similar spin density using SIESTA. The generalized gradient approximation (GGA) of the exchange and correlation functional was used with the Perdew-Burke-Ernzerhof parameterization, a double- ζ basis set, a real-space grid defined with an equivalent energy cut-off of 150 Ry and a force tolerance of less than 10 meV Å⁻¹. For the band structure calculation, each structure was sampled by a $1 \times 1 \times 15$ Monkhorst-Pack k -point grid. We found the stable magnetic state by allowing the system to be spin-polarized. Apart from the two edge atoms on the peripheral NIT fivefold rings, atoms were numbered such that odd-numbered atoms were connected only to odd-numbered atoms and the same for even-numbered ones. We then performed geometry optimization of each system by choosing the initial system to have a ferromagnetic configuration (all spin-up), or an antiferromagnetic configuration in which the odd (even) atoms were designated spin-up (spin-down). The total energy per unit cell of the antiferromagnetically aligned NIT-GNRs (NIT-polyphenylene) is 272 meV (651 meV) lower than that of the ferromagnetically aligned ones. The molecular orbitals, band structure and local density of state calculations were obtained using the antiferromagnetic spin alignment.

Data availability. The datasets generated and analysed during this study are available from the corresponding author on reasonable request.

- Narita, A. et al. Bottom-up synthesis of liquid-phase-processable graphene nanoribbons with near-infrared absorption. *ACS Nano* **8**, 11622–11630 (2014).
- Keerthi, A. et al. Hexa-peri-hexabenzocoronene with different acceptor units for tuning optoelectronic properties. *Chem. Asian J.* **11**, 2710–2714 (2016).
- Stoll, S. & Schweiger, A. EasySpin, a comprehensive software package for spectral simulation and analysis in EPR. *J. Magn. Reson.* **178**, 42–55 (2006).
- Wacker, T., Sierra, G. A. & Schweiger, A. The concept of FID-detected hole-burning in pulsed EPR spectroscopy. *Isr. J. Chem.* **32**, 305–322 (1992).
- Rintoul, L., Micallef, A. S. & Bottle, S. E. The vibrational group frequency of the N-O stretching band of nitroxide stable free radical. *Spectrochim. Acta A* **70**, 713–717 (2008).
- Cox, N., Lubitz, W. & Savitzky, A. W-band ELDOR-detected NMR (EDNMR) spectroscopy as a versatile technique for the characterisation of transition metal-ligand interactions. *Mol. Phys.* **111**, 2788–2808 (2013).
- Jeschke, G. et al. DeerAnalysis2006—a comprehensive software package for analyzing pulsed ELDOR data. *Appl. Magn. Reson.* **30**, 473–498 (2006).
- Gaussian 09, revision A.02. (Gaussian, Wallingford, 2016).
- Soler, J. M. et al. The SIESTA method for *ab initio* order- N materials simulation. *J. Phys. Condens. Matter* **14**, 2745–2779 (2002).

Approaching the Schottky–Mott limit in van der Waals metal–semiconductor junctions

Yuan Liu^{1,2}, Jian Guo¹, Enbo Zhu¹, Lei Liao², Sung-Joon Lee¹, Mengning Ding¹, Imran Shakir³, Vincent Gambin⁴, Yu Huang^{1,5*} & Xiangfeng Duan^{5,6*}

The junctions formed at the contact between metallic electrodes and semiconductor materials are crucial components of electronic and optoelectronic devices¹. Metal–semiconductor junctions are characterized by an energy barrier known as the Schottky barrier, whose height can, in the ideal case, be predicted by the Schottky–Mott rule^{2–4} on the basis of the relative alignment of energy levels. Such ideal physics has rarely been experimentally realized, however, because of the inevitable chemical disorder and Fermi-level pinning at typical metal–semiconductor interfaces^{2,5–12}. Here we report the creation of van der Waals metal–semiconductor junctions in which atomically flat metal thin films are laminated onto two-dimensional semiconductors without direct chemical bonding, creating an interface that is essentially free from chemical disorder and Fermi-level pinning. The Schottky barrier height, which approaches the Schottky–Mott limit, is dictated by the work function of the metal and is thus highly tunable. By transferring metal films (silver or platinum) with a work function that matches the conduction band or valence band edges of molybdenum sulfide, we achieve transistors with a two-terminal electron mobility at room temperature of 260 centimetres squared per volt per second and a hole mobility of 175 centimetres squared per volt per second. Furthermore, by using asymmetric contact pairs with different work functions, we demonstrate a silver/molybdenum sulfide/platinum photodiode with an open-circuit voltage of 1.02 volts. Our study not only experimentally validates the fundamental limit of ideal metal–semiconductor junctions but also defines a highly efficient and damage-free strategy for metal integration that could be used in high-performance electronics and optoelectronics.

Metal–semiconductor junctions are at the heart of modern electronics and optoelectronics. One of the most important parameters for the metal–semiconductor junction is the Schottky barrier height (Φ_{SB}), an energy barrier for a charge carrier to cross the junction, which can fundamentally determine charge transport efficiency and impact device performance^{1,2}. In an ideal metal–semiconductor junction, Φ_{SB} can be well predicted by the Schottky–Mott rule, a law first proposed in the 1930s and governed by electrostatics in all types of problem that involve energy-level alignments^{3,4}:

$$\Phi_{\text{SB},n} = \Phi_{\text{M}} - X_{\text{S}} \quad (1)$$

$$\Phi_{\text{SB},p} = I_{\text{S}} - \Phi_{\text{M}} \quad (2)$$

where Φ_{M} is the work function of the metal, X_{S} and I_{S} are the electron affinity and ionization potential of the semiconductor, correspondingly, and $\Phi_{\text{SB},n}$ and $\Phi_{\text{SB},p}$ are the Schottky barrier heights for electrons and holes, respectively. These quantities are the intrinsic properties of the isolated materials before they form the junction, and the Schottky–Mott model implies that Φ_{SB} is linearly dependent on the metal work function with a slope of unity.

Experimentally, however, the Schottky–Mott model gives grossly incorrect predictions for the Schottky barrier height²: Φ_{SB} is usually insensitive to Φ_{M} , and the Fermi level of the system is typically pinned to a nearly fixed position within the semiconductor bandgap, varying little with respect to different metals used, as first noted by Bardeen in 1947 (ref. ⁵). The strength of Fermi-level pinning (FLP) for a given semiconductor can be characterized by the interface S parameter:

$$S = |d\Phi_{\text{SB}}/d\Phi_{\text{M}}| \quad (3)$$

If $S = 1$, the Schottky–Mott limit is achieved. Unfortunately, S is generally far less than unity for most typical semiconductors (approximately 0.27 for Si and 0.07 for GaAs)¹³ and the Schottky–Mott limit has not been experimentally achieved in traditional metal–semiconductor junctions.

This striking discrepancy between theory and experiment arises because the Schottky–Mott model is purely dependent on ideal physics and neglects several types of chemical interaction that are hard to avoid at the interface of two dissimilar materials. First, owing to the termination of the crystal structure and incomplete covalent bonds, surface dangling bonds or surface reconstructions lead to surface states (Bardeen pinning effect or Shockley–Tamm states) within the semiconductor bandgap and result in FLP at these energy levels⁵. Second, the interface of the contact is rarely an atomically sharp discontinuity between the metal and the semiconductor crystal, where chemical bonds take place and modify their original energy levels. Chemical bonding between the metal and the semiconductor, and their interdiffusion, can also create large strain in both crystal lattices and change the band structures, as well as the resulting barrier height^{6–8}. Third, the typical processes for material integration and device fabrication usually lead to additional chemical disorders and defect-induced gap states that serve as a reservoir for electrons or holes and therefore pin the Fermi level⁹. For example, ‘high-energy’ metal deposition processes usually involve atom or cluster bombardment and strong local heating to the contact region, which could damage the crystal lattice at or near the interface¹⁰, as commonly observed in III–V compound semiconductors¹¹; moreover, the resist development process could also leave polymer residue within the interface that causes the overall measured barrier height to deviate from the predicted value. Fourth, metal-induced gap states are formed in the junction owing to the decaying metallic wavefunction that penetrate to nanometre depth into the semiconductor¹².

Here we report the design and creation of van der Waals (vdW) metal–semiconductor junctions, in which metal electrodes with atomically flat surfaces are pre-fabricated and physically laminated¹⁴ onto dangling-bond-free two-dimensional (2D) semiconductors without direct chemical bonding, avoiding the associated chemical disorder and defect-induced gap states. The fabrication process flow is illustrated in Fig. 1a. Briefly, a series of metal electrodes with various work functions are first prepared on a silicon substrate. They can be

¹Department of Materials Science and Engineering, University of California, Los Angeles, CA, USA. ²State Key Laboratory for Chemo/Biosensing and Chemometrics, College of Chemistry and Chemical Engineering, and School of Physics and Electronics, Hunan University, Changsha, China. ³Sustainable Energy Technologies Centre, College of Engineering, King Saud University, Riyadh, Saudi Arabia. ⁴NG NEXT, Northrop Grumman Corporation, Redondo Beach, CA, USA. ⁵California Nanosystems Institute, University of California, Los Angeles, CA, USA. ⁶Department of Chemistry and Biochemistry, University of California, Los Angeles, CA, USA. *e-mail: yhuang@seas.ucla.edu; xduan@chem.ucla.edu

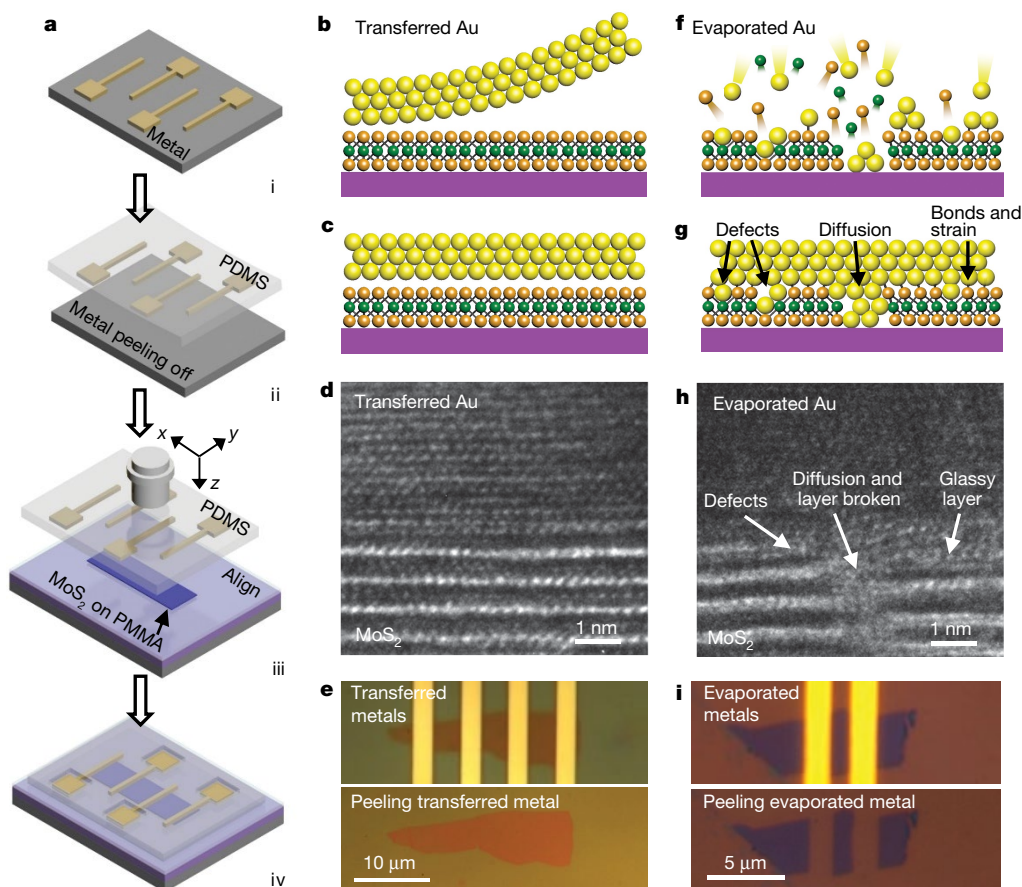


Fig. 1 | Illustration and structural characterizations of vdW metal-semiconductor junctions. **a**, Schematic illustrations of vdW integration of metal-semiconductor junctions: (i) metal deposition on sacrificial substrate; (ii) peeling off the metal; (iii) alignment; and (iv) contact lamination and probe window opening. **b–d**, Cross-sectional schematics and TEM images of the transferred Au electrode on top of MoS₂, with atomically sharp and clean metal-semiconductor interfaces. **e**, Optical image of the MoS₂ device with transferred electrodes (upper) and with the transferred electrodes mechanically released (lower). The underlying MoS₂ layer retains its original shape after physical integration and separation of the Au thin-film electrodes, indicating that the transferred

mechanically released from the silicon and display an atomically flat surface, replicating the flat surface of the substrate (see Methods and Extended Data Fig. 1 for details). Next, few-layer MoS₂ flakes (4–20 nm, unless otherwise specified) are mechanically exfoliated on top of highly doped silicon (p⁺⁺) covered with SiO₂ (300 nm) and poly(methyl methacrylate) (PMMA, 170 nm) as dielectrics (Fig. 1a, ii). PMMA here functions as a dielectric that is nearly free from trap states^{15,16} and is essential for retaining the nearly intrinsic electronic properties of MoS₂, in contrast to a conventional MoS₂/SiO₂ interface with rich trap states (Extended Data Fig. 2). Next, the previously released metal electrodes are aligned under a microscope and physically laminated on top of the MoS₂ flake, resulting in an atomically sharp and clean metal-semiconductor interface (Fig. 1a, iii). Finally, the PMMA on top of the contact pads is removed using standard electron-beam lithography and development processes, leaving the exposed metal pads for electrical probing and measurements (Fig. 1a, iv).

This vdW integration of metal thin-film electrodes and 2D semiconductors has several advantages that could overcome the typical FLP limitations and lead to an interface approaching the ideal physical model. First, in terms of semiconductor surface, the dangling-bond-free surface of the 2D semiconductors could eliminate the Shockley–Tamm states that dominate a three-dimensional covalent semiconductor surface with rich surface dangling bonds or surface reconstructions^{17,18}. Second, the physical transfer of pre-fabricated metal electrodes offers

metal-semiconductor interface is free of direct chemical bonding. **f–h**, Cross-sectional schematics and TEM images of conventional electron-beam-deposited Au electrodes on top of MoS₂, where the bombardment of the MoS₂ surface by high-energy Au atoms and clusters creates considerable damage to the MoS₂ surface, producing a glassy layer with apparent defects, interface diffusion, chemical bonding and atomic disorder. **i**, Optical image of a MoS₂ flake with deposited electrodes (upper) and with the deposited electrodes mechanically released (lower), where the underlying MoS₂ surface is destroyed while removing the deposited electrodes, suggesting direct chemical bonding and strong metal-semiconductor interaction in deposited junctions.

a gentle, ‘low-energy’ materials integration strategy without conventional aggressive fabrication processes (for example, lithography or deposition) to prevent the creation of defects, residues, strains and the associated defect-induced gap states on a dangling-bond-free 2D semiconductor surface. This can be clearly seen in cross-sectional transmission electron microscopy (TEM) images, in which the transferred metal/MoS₂ junctions feature an atomically sharp and clean interface (Fig. 1b–d), whereas the deposited metal/MoS₂ interfaces show considerable defects, strain, disorder and metal diffusion (Fig. 1f–h). Third, the physical contact without direct chemical bonding can greatly suppress the interface dipoles and metal-induced gap states^{19–22}.

To demonstrate the weak vdW interaction at the interface, we mechanically separated the transferred metal electrodes from MoS₂ after the device fabrication and electrical measurement. The underlying semiconductor retains its original shape without any apparent damage (Fig. 1e). In contrast, the deposited metal electrodes typically form strong chemical bonding with the underlying MoS₂ (such as Au–S bonds), generating a glassy layer dominated by interdiffusion and strain. When the deposited metal electrodes are mechanically peeled²³, the underlying MoS₂ is destroyed at the same time (Fig. 1i). The reversible physical integration and isolation of the transferred metal-semiconductor junctions are strong indicators of ideal interfaces, where two crystals in intimate contact retain their isolated states without direct chemical bonding.

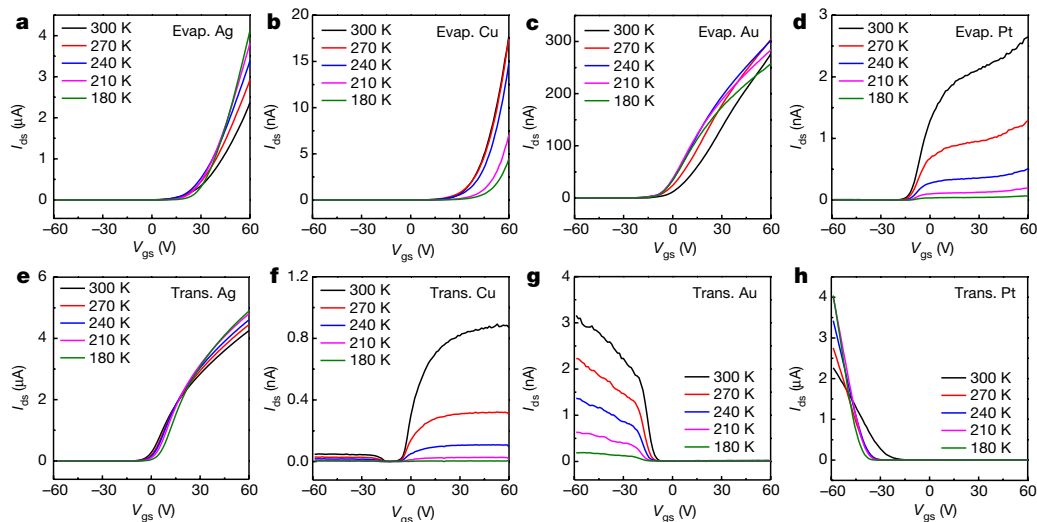


Fig. 2 | Transfer characteristics of MoS₂ transistors with deposited and transferred metal electrodes. **a–d**, I_{ds} – V_{gs} transfer curves of MoS₂ transistors with Ag (**a**), Cu (**b**), Au (**c**) and Pt (**d**) electrodes deposited by electron-beam evaporation. We always observe n-type behaviour irrespective of the highly distinct work function of the contact metal used, which suggests strong Fermi-level pinning near the conduction band edge. **e–h**, I_{ds} – V_{gs} transfer curves of MoS₂ transistors with transferred Ag (**e**),

Cu (**f**), Au (**g**) and Pt (**h**) electrodes. The device switches from n-type to p-type characteristics with increasing work function of the contact electrodes, suggesting highly tunable electron and hole barriers depending on the work function of the transferred contact metal used. The bias voltage is 100 mV, and the gate dielectric is composed of 300-nm-thick SiO₂ and 170-nm-thick PMMA for all measurements. Evap., evaporated; trans., transferred.

With minimized interface disorder and weak metal–semiconductor interaction, the vdW-contacted MoS₂ transistors exhibit highly tunable device characteristics dictated by the metal work functions. Figure 2 shows the I_{ds} – V_{gs} transfer curve of MoS₂ transistors contacted by a series of transferred metals with various work functions. For comparison, we also characterized the devices using the same metal electrodes but prepared using conventional electron-beam evaporation. In general, the MoS₂ devices contacted by conventional evaporation-deposited metals show n-type behaviour regardless of the work function of the metal used (Fig. 2a–d), which is consistent with previous studies²⁴ and strongly indicates that the Fermi level is pinned near the conduction band edge of MoS₂.

By contrast, for devices with transferred metal electrodes, the majority carrier type can be systematically tailored from electrons to holes by varying the work function of the contact metals (Fig. 2e–h). For example, for Ag, which has a low work function ($W_{Ag} \approx 4.3$ eV), well behaved n-type transfer curves are observed (Fig. 2e) with a typical metal–insulator transition behaviour^{25,26}, indicating optimized metal–semiconductor contact and low electron barrier. Next, for Cu, which has a medium work function ($W_{Cu} \approx 4.6$ eV), the device exhibits a bipolar transfer curve with preferred n-type behaviour (Fig. 2f). Compared with Ag-contacted devices, the I_{ds} current is three orders of magnitude smaller at room temperature and decreases exponentially with temperature, demonstrating a relatively large n-type Schottky barrier. On further increasing the metal work function by using an Au electrode ($W_{Au} \approx 5.1$ eV), the device exhibits predominantly p-type behaviour with a low current level (nanoamperes) (Fig. 2g). The I_{ds} drops quickly with decreasing temperature, suggesting that a large p-type Schottky barrier dominates the overall carrier transport. This is in contrast to previous devices (and control samples in Fig. 2c) with deposited Au electrodes as n-type ohmic contacts to MoS₂. In those devices, formation of Au–S bonding²⁷ dominates the carrier transport with strong FLP near the conduction band edge (Fig. 2c). Finally, for transferred Pt, with the highest work function ($W_{Pt} \approx 5.6$ eV), the device shows well behaved p-type characteristics (Fig. 2h) with an ohmic I_{ds} – V_{ds} output curve (Extended Data Fig. 3). When reducing the temperature, p-type metal–insulator transition is observed in MoS₂, suggesting an optimized p-type contact with negligible hole barrier. In contrast, the device with deposited Pt electrodes exhibits poor n-type behaviour due to FLP near the conduction band edge (Fig. 2d), consistent with previous studies²⁴.

Our results above demonstrate that carrier transport in MoS₂ transistors can be systematically tailored by using transferred metal contacts with different work functions. To further evaluate the dependence on different metals, we have extracted the Schottky barrier height using the equation:

$$I_{ds} = AA^*T^2 \exp\left[-\frac{\Phi_{SB}}{kT}\right] \quad (4)$$

where I_{ds} is the current through the device, A is the junction area, A^* is the Richardson constant, k is the Boltzmann constant and T is the temperature. We note that Φ_{SB} here is extracted under the flat-band condition (see Methods), in which the tunnelling current across the Schottky barrier can be minimized^{24,28}; detailed description of the extraction can be found in Methods and Extended Data Fig. 4. Figure 3 shows the extracted Schottky barrier height for different metals used in our study as a function of the corresponding work functions; the solid line is the linear fitting of the results, the slope of which corresponds to the interface S parameter. For control devices with deposited metals, the extracted S parameter is 0.09, consistent with previous studies²⁴ of MoS₂ with $S = 0.1$, confirming strong FLP near the conduction band edge at the metal/MoS₂ interface (largely due to fabrication-induced defects and gap states; see Fig. 1f–h). By contrast, for the devices with transferred metal electrodes, the Φ_{SB} value is strongly dependent on the metal work functions, and the Schottky barrier type can be tuned from electrons to holes. The fitted S parameter is 0.96, approaching the limit of the Schottky–Mott law defined by electrostatic energy alignment. S is also much larger than the previously reported values^{1,6} of 0.27 for Si and 0.07 for GaAs, indicating a nearly ideal interface between the physically transferred metal contact and the dangling-bond-free 2D semiconductor surface, in contrast to the inevitable chemical disorder and FLP at typical metal–semiconductor interfaces fabricated previously.

The ability to prepare atomically sharp and atomically clean metal–semiconductor interfaces and to tailor the Schottky barrier height opens a pathway to overcome the FLP effect that plagues 2D semiconductor devices and to improve their performance. For instance, by applying transferred Ag electrodes with a low electron barrier, we have fabricated an n-channel MoS₂ transistor with two-terminal electron mobility (μ_e) reaching $260 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$, considerably higher than previous reports for two-terminal back-gated MoS₂ devices²⁴ (Extended

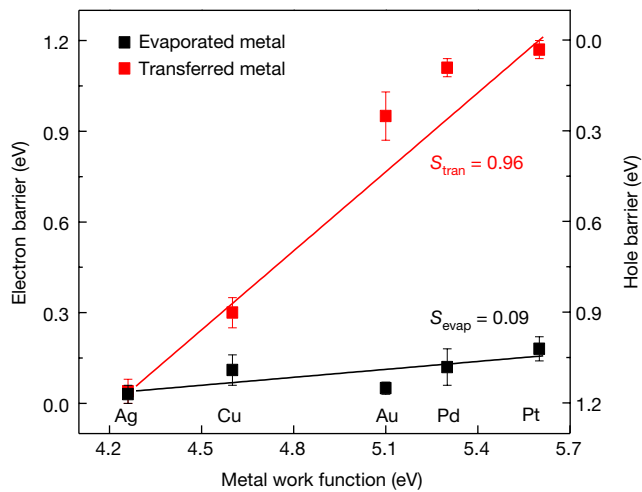


Fig. 3 | Experimentally determined Schottky barrier height for different transferred metals and evaporated metals. For transferred metal electrodes, the majority carrier type and corresponding Schottky barrier height is strongly dependent on the metal work function with a slope ($S = 0.96$) approaching unity, suggesting excellent obedience to the Schottky–Mott law. With the conventional evaporation-deposited metal electrodes, the devices invariably show n-type behaviour with a small electron Schottky barrier and a slope $S = 0.09$, indicating the strong pinning effect at the metal–semiconductor interface.

Data Fig. 5 and Extended Data Table 1) that did not exclude contact resistance. By shortening the channel length to 160 nm, we can further increase the delivering current density to $0.66 \text{ mA } \mu\text{m}^{-1}$ (Extended Data Fig. 6) at room temperature, comparable to the best reported devices using vdW graphene hybrid contacts or strong contact doping

(Extended Data Table 1). Similarly, by using transferred Pt contacts with minimized barrier for holes in MoS₂, we have achieved higher two-terminal hole mobility (μ_h) of $175 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ and hole current density of $0.21 \text{ mA } \mu\text{m}^{-1}$ in p-channel MoS₂ transistors than in previous work (Extended Data Figs. 3 and 6, and Extended Data Table 1). It should be noted that although the non-bonding vdW gap may pose an additional series tunnelling resistance, its value is too small to affect the overall carrier transport (see Methods) and can be largely neglected.

Taking the method a step further, we can also transfer pairs of metals with distinct work functions to enable high-performance optoelectronic devices beyond the reach of the traditional processes. For example, we have created a metal–semiconductor–metal (MSM) photodiode by using a transferred Ag and Pt electrode pair as vdW contacts (Fig. 4a). With the asymmetric Ag and Pt contacts, the device shows rectification behaviour, with a rectification ratio of up to 10^8 and an ideality factor η of 1.09 (Fig. 4b and Extended Data Fig. 7). The near-unity ideality factor obtained in the transferred MSM diode is much better than that of the deposited MSM device with $\eta > 1.8$ (Extended Data Fig. 8), confirming the high interfacial quality of the vdW metal–semiconductor junctions.

Under wide-field laser illumination (532 nm , $10 \text{ nW } \mu\text{m}^{-2}$), our Ag–MoS₂–Pt MSM photodiodes with transferred contacts produce a remarkable open-circuit voltage V_{oc} of 1.02 V in the monolayer MoS₂ (bandgap $E_g \approx 1.8 \text{ V}$) device and 0.76 V in the seven-layer MoS₂ ($E_g \approx 1.2 \text{ V}$) device (Fig. 4c and Extended Data Fig. 7). The V_{oc} value of 0.76 V is more than twice that of a control device with deposited Pt–Ag contacts (Extended Data Fig. 8). Overall, the V_{oc} achieved in the MSM diode with transferred Ag and Pt contacts is considerably higher than those of 2D semiconductor MSM or p–n photodiodes reported previously ($0.1\text{--}0.8 \text{ V}$) (Extended Data Table 2). The lower V_{oc} obtained in previous 2D semiconductor photodiodes may be partly attributed to the difficulties in achieving a low contact barrier for both

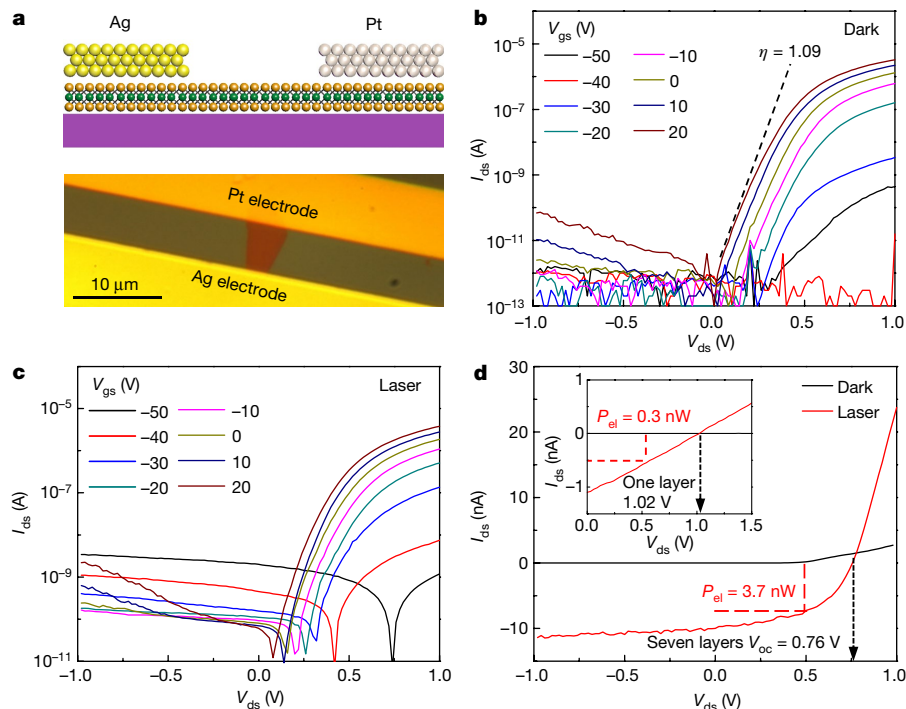


Fig. 4 | The Ag–MoS₂–Pt MSM photodiode with transferred asymmetric Ag–Pt electrodes. **a**, Schematic illustration and optical image of transferred asymmetric Ag and Pt electrodes on MoS₂. The Ag is grounded, and the Pt electrode is used as the drain. **b**, **c**, Semi-logarithmic plot of the I_{ds} – V_{ds} output curve of a seven-layer device under a dark environment (**b**) and under 532-nm laser illumination (**c**, $10 \text{ nW } \mu\text{m}^{-2}$). The diode demonstrates a high rectification ratio ($> 10^8$), near-unity ideality factor ($\eta = 1.09$) and a large V_{oc} . **d**, Linear plot of I_{ds} – V_{ds}

output curve under dark (black line) and laser illumination (red line), demonstrating a highest V_{oc} of 1.02 V for monolayer MoS₂ (inset) and 0.76 V for seven-layer MoS₂. The gate voltage is -60 V for the monolayer device and -50 V for the seven-layer device. The red dashed lines show the corresponding power area for maximum power conversion. The gate dielectric is composed of 300-nm -thick SiO₂ and 170-nm -thick PMMA for all devices in optoelectronic measurement. P_{el} , electrical power.

electrons and holes in the same device. By contrast, with the optimized asymmetric vdW contacts in the Ag–MoS₂–Pt MSM photodiodes, the contact barrier for electrons is minimized at the Ag–MoS₂ interface, and the contact barrier for holes is minimized at the MoS₂–Pt interface, thus ensuring a high V_{oc} . The Ag–MoS₂–Pt photodiodes give a photoresponsivity of 7.2 mA W^{−1} and 16.6 mA W^{−1}, and external quantum efficiencies of 1.74% and 4.5% for the monolayer and seven-layer devices, respectively, higher than those of previous p–n junctions (about 0.2%)^{29,30} made by dual-gated WSe₂. A maximum electrical power output of 0.3 nW (3.7 nW) is obtained at $V_m = 0.54$ V (0.5 V) for the monolayer (multilayer) device (Fig. 4d), with a power conversion efficiency of 0.2% (0.6%) (see Methods).

Our study not only validates the fundamental limit of ideal metal–semiconductor interfaces, but also defines a general, low-energy metal integration approach that may be extended to other delicate materials that would be damaged by aggressive contact fabrication processes (for example, degradation in various solvents used in lithography processes or in ‘high-energy’ metal deposition) or to other functional interfaces or junctions (for example, magnetic–semiconductor or superconductor–semiconductor junctions) that were previously limited by interface disorder.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0129-8>.

Received: 27 November 2017; Accepted: 19 February 2018;

Published online: 16 May 2018

- Sze, S. M. & Ng, K. K. *Physics of Semiconductor Devices* (Wiley, Hoboken, 2006).
- Tung, R. T. The physics and chemistry of the Schottky barrier height. *Appl. Phys. Rev.* **1**, 011304 (2014).
- Schottky, W. Zur Halbleitertheorie der Sperrschicht- und Spitzengleichrichter. *Z. Phys. A* **113**, 367 (1939).
- Mott, N. The theory of crystal rectifiers. *Proc. R. Soc. Lond. A* **171**, 27–38 (1939).
- Bardeen, J. Surface states and rectification at a metal semi-conductor contact. *Phys. Rev.* **71**, 717–727 (1947).
- Saidi, W. A. Influence of strain and metal thickness on metal–MoS₂ contacts. *J. Chem. Phys.* **141**, 094707 (2014).
- Kang, J., Liu, W., Sarkar, D., Jena, D. & Banerjee, K. Computational study of metal contacts to monolayer transition-metal dichalcogenide semiconductors. *Phys. Rev. X* **4**, 031005 (2014).
- Tung, R. T. Chemical bonding and Fermi level pinning at metal–semiconductor interfaces. *Phys. Rev. Lett.* **84**, 6078–6081 (2000).
- Hasegawa, H. & Sawada, T. On the electrical properties of compound semiconductor interfaces in metal/insulator/semiconductor structures and the possible origin of interface states. *Thin Solid Films* **103**, 119–140 (1983).
- Zan, R. et al. Control of radiation damage in MoS₂ by graphene encapsulation. *ACS Nano* **7**, 10167–10174 (2013).
- Spicer, W., Chye, P., Garner, C., Lindau, I. & Pianetta, P. The surface electronic structure of 3–5 compounds and the mechanism of Fermi level pinning by oxygen (passivation) and metals (Schottky barriers). *Surf. Sci.* **86**, 763–788 (1979).
- Heine, V. Theory of surface states. *Phys. Rev.* **138**, A1689–A1696 (1965).
- Cowley, A. & Sze, S. Surface states and barrier height of metal–semiconductor systems. *J. Appl. Phys.* **36**, 3212–3220 (1965).
- Loo, Y.-L. et al. Soft, conformable electrical contacts for organic semiconductors: high-resolution plastic circuits by lamination. *Proc. Natl Acad. Sci. USA* **99**, 10252–10256 (2002).
- Bao, W., Cai, X., Kim, D., Sridhara, K. & Fuhrer, M. S. High mobility ambipolar MoS₂ field-effect transistors: substrate and dielectric effects. *Appl. Phys. Lett.* **102**, 042104 (2013).
- Liu, Y. et al. Vertical charge transport and negative transconductance in multilayer molybdenum disulfides. *Nano Lett.* **17**, 5495–5501 (2017).
- Liu, Y. et al. Van der Waals heterostructures and devices. *Nat. Rev. Mater.* **1**, 16042 (2016).
- Novoselov, K., Mishchenko, A., Carvalho, A. & Neto, A. C. 2D materials and van der Waals heterostructures. *Science* **353**, aac9439 (2016).
- Gong, C., Colombo, L., Wallace, R. M. & Cho, K. The unusual mechanism of partial Fermi level pinning at metal–MoS₂ interfaces. *Nano Lett.* **14**, 1714–1720 (2014).
- Liu, Y., Stradins, P. & Wei, S.-H. Van der Waals metal–semiconductor junction: weak Fermi level pinning enables effective tuning of Schottky barrier. *Sci. Adv.* **2**, e1600069 (2016).
- Farmanbar, M. & Brocks, G. Controlling the Schottky barrier at MoS₂/metal contacts by inserting a BN monolayer. *Phys. Rev. B* **91**, 161304 (2015).
- Farmanbar, M. & Brocks, G. First-principles study of van der Waals interactions and lattice mismatch at MoS₂/metal interfaces. *Phys. Rev. B* **93**, 085304 (2016).
- Desai, S. B. et al. Gold mediated exfoliation of ultralarge optoelectronically perfect monolayers. *Adv. Mater.* **28**, 4053–4058 (2016).
- Das, S., Chen, H.-Y., Penumatcha, A. V. & Appenzeller, J. High performance multilayer MoS₂ transistors with scandium contacts. *Nano Lett.* **13**, 100–105 (2013).
- Cui, X. et al. Multi-terminal transport measurements of MoS₂ using a van der Waals heterostructure device platform. *Nat. Nanotech.* **10**, 534–540 (2015).
- Liu, Y. et al. Toward barrier free contact to molybdenum disulfide using graphene electrodes. *Nano Lett.* **15**, 3030–3034 (2015).
- Popov, I., Seifert, G. & Tománek, D. Designing electrical contacts to MoS₂ monolayers: a computational study. *Phys. Rev. Lett.* **108**, 156802 (2012).
- Cui, X. et al. Low-temperature ohmic contact to monolayer MoS₂ by van der Waals bonded Co/h-BN electrodes. *Nano Lett.* **17**, 4781–4786 (2017).
- Pospischil, A., Furchi, M. M. & Mueller, T. Solar-energy conversion and light emission in an atomic monolayer PN diode. *Nat. Nanotech.* **9**, 257–261 (2014).
- Baughner, B. W., Churchill, H. O., Yang, Y. & Jarillo-Herrero, P. Optoelectronic devices based on electrically tunable pn diodes in a monolayer dichalcogenide. *Nat. Nanotech.* **9**, 262–267 (2014).

Acknowledgements X.D. acknowledges support by the Office of Naval Research through grant number N00014-15-1-2368. Y.H. acknowledges support by the National Science Foundation EFRI-1433541. I.S. thanks the Deanship of Scientific Research at King Saud University for its funding of this research through grant number PEJP-17-01. L.L. acknowledges support by the National Key Research and Development Program of China number 2016YFB0401103. We acknowledge the Electron Imaging Center at UCLA for TEM technical support and the Nanoelectronics Research Facility at UCLA for device fabrication technical support.

Author contributions X.D. and Y.H. conceived the research. X.D. and Y.L. designed the experiments. Y.L. performed the experiments and data analysis. J.G., S.L. and M.D. contributed to the device fabrication. E.Z. contributed to cross-sectional TEM characterization. L.L., I.S. and V.G. contributed to discussions and data analysis. X.D. and Y.L. co-wrote the manuscript. All authors discussed the results and commented on the manuscript.

Competing interests The authors declare no competing financial interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0129-8>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to Y.H. or X.D. **Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Metal electrode fabrication, release, transferring and lamination process. We first prepare a series of 50-nm-thick metal electrodes with various work functions on a silicon substrate with an atomically flat surface by using standard photolithography or electron-beam lithography and high-vacuum electron-beam evaporation. Next, a hexamethyldisilazane (HMDS) layer is applied to functionalize the whole wafer, and a PMMA layer is then spin-coated on top of the metal electrodes. With the pre-functionalization by HMDS, the PMMA layer has weak adhesion to the sacrificial substrate and can be mechanically released by using a PDMS stamp, together with metal electrodes wrapped underneath (Fig. 1a, ii). The metal electrodes released using this method are atomically flat (replicating the atomically flat surface of the sacrificial wafer), with a mean surface roughness of 0.2–0.3 nm (Extended Data Fig. 1).

In the functionalization step, we place the wafer and metal electrodes in a sealed HMDS chamber at 120 °C for 2–30 min and then spin-coat it with the PMMA polymer. Both the required HMDS functionalization time and the PMMA thickness are highly dependent on the metal to be released. For metals with weak adhesion (Au, Ag, Pt) to the silicon wafer, the functionalization time is 20–30 min and the PMMA polymer is about 1 µm thick. For metals with intermediate adhesion force (Pd), the treatment time is 5 min and the PMMA polymer is about 2 µm thick. For metals with greater adhesion force (Cu), the treatment time is 3 min and the PMMA polymer is about 5–10 µm thick. For metals with the strongest adhesion force (Ti, Ni, Cr), the treatment time is <2 min and the PMMA polymer is >10 µm thick. However, for Ti, Ni and Cr, the releasing yield is low, and this release method needs further improvement. Extended Data Fig. 1 shows optical images and photographs taken during the transfer process. For Au electrode transfer, the highest transfer temperature (from PDMS to the target substrate) should be kept lower than 60 °C, to maintain the metal–semiconductor vdW gap and to avoid their strong interaction and formation of chemical bonds. The transfer process is conducted in a nitrogen-filled glovebox with low oxygen level (<0.1 p.p.m.). Furthermore, once the metal is delaminated from the substrate, it is physically contacted onto the MoS₂ immediately with short exposure time (typically <20 s) to minimize any possible surface oxide formation on oxygen-sensitive metals (for example, Cu and Ag).

Flat-band Schottky barrier extraction. Our extraction of the Schottky barrier height is based on a thermionic model under low doping level, as shown in Extended Data Fig. 4. At low doping level, below the flat-band voltage (V_{FB}), the charge injection to the MoS₂ channel is mainly through thermionic emission with the relationship in equation (4). In this way, the Schottky barrier Φ_{SB} can be extracted by using Arrhenius plots with the following equation^{26,31}:

$$\ln\left(\frac{I_{ds}}{T^2}\right) = -\frac{\Phi_{SB}}{kT} + c \quad (5)$$

where c is a constant and Φ_{SB} is the slope (between $-1/kT$ and $\ln(I_{ds}/T^2)$) within the Arrhenius plots. Using equations (4) and (5), we can extract the Schottky barrier height under various gate voltages, as summarized in Extended Data Fig. 4c, f.

To accurately represent the thermionic emission of the metal–semiconductor interface, the Schottky barrier under a flat-band gate voltage ($V_{gs} = V_{FB}$) is always used in Fig. 3. Below the flat-band voltage ($V_{gs} < V_{FB}$), the device is in the subthreshold regime and the channel resistance dominates the carrier transport. Above the flat-band voltage ($V_{gs} > V_{FB}$), the contact is highly doped, and a superimposed tunnelling current affects the extracted barrier height, resulting in apparently smaller Φ_{SB} . In theory, the extracted Schottky barrier Φ_{SB} has a linear relationship with the gate voltage in the subthreshold regime and gradually becomes sublinear above V_{FB} (refs^{28,32}). In this way, we could use a guided line to extract the V_{FB} value and accurately determine the flat-band Schottky barrier (Extended Data Fig. 4c, f).

For the extraction of the Schottky barrier height, a 0.1 V bias voltage (V_{ds}) is used. The resulting relationship between I_{ds} and the diode saturation current can be expressed using the typical back-to-back Schottky diode model with a simple current continuity equation:

$$I_F = I_{sat} \left(e^{\frac{qV_F}{kT}} - 1 \right) \quad (6)$$

$$I_R = I_{sat} \left(e^{\frac{-qV_R}{kT}} - 1 \right) \quad (7)$$

$$I_{ds} = I_F = -I_R \quad (8)$$

$$V_{ds} = V_F + V_R \quad (9)$$

where I_F is the current of the forward-bias diode, V_F is the voltage applied to the forward-bias diode, I_R is the current of the reverse-bias diode, V_R is the voltage

applied to the reverse-bias diode and I_{sat} is the saturation current that we need to measure. On the basis of equations (6)–(9) with $V_{ds} = 0.1$ V, the overall measured current I_{ds} is 96% of I_{sat} , approaching the case for a nearly ideal single diode; this indicates the accurate measurement of I_{sat} under this condition. Therefore, the relative small bias voltage used here not only can minimize the superimposed tunnelling current to approach an ideal Schottky barrier based on pure thermionic emission but is also large enough for the whole system to be viewed as a single diode at the source side.

Impact of ultra-thin vdW gap for carrier transport. Although the non-bonding vdW gap may pose an additional series resistance, its value is too small (0.1 nm to 0.15 nm) to affect the overall carrier transport and can be largely neglected. To quantitatively verify this argument, we have calculated the tunnelling resistance of a vdW gap within the metal–semiconductor interface, through a direct tunnelling model with the following relationship³³:

$$J_T = \frac{q^3}{16\pi^2\hbar\varphi_b} F_{vdW}^2 \times \exp\left\{-\frac{4}{3} \frac{\sqrt{2m_{vdW}} \varphi_b}{\hbar q F_{vdW}} \left[1 - \left(1 - \frac{qV_{vdW}}{\varphi_b}\right)^{\frac{3}{2}}\right]\right\} \quad (10)$$

where J_T is the calculated current density, q is the electron charge, \hbar is the reduced Planck constant, φ_b is the distance between the MoS₂ conduction band and the vacuum energy, F_{vdW} is the electrical field within the vdW gap, m_{vdW} is the electron mass within the vdW gap and V_{vdW} is the bias voltage applied across the vdW gap.

As shown in equation (10), to accurately evaluate the tunnelling resistance, we have further determined the vdW gap thickness using both theoretical calculations and experimental measurements. In theory, the vdW gap thickness (T_{theory}) between metal and MoS₂ can be calculated by subtracting the atomic radius from its vdW diameter, using the following equations:

$$T_{theory} = (r_{vdW(m)} - r_{atom(m)}) + (r_{vdW(s)} - r_{atom(s)}) \quad (11)$$

where $r_{vdW(s)}$, $r_{vdW(m)}$, $r_{atom(s)}$, $r_{atom(m)}$ are the vdW radius of sulfur, vdW radius of the metal, atomic radius of sulfur and atomic radius of the metal, respectively. Given the r_{atom} and r_{vdW} values are 0.88 Å and 1.8 Å for sulfur; 1.35 Å and 1.4 Å for Cu; 1.4 Å and 1.63 Å for Pd; 1.6 Å and 1.72 Å for Ag; 1.35 Å and 1.66 Å for Au; and 1.35 Å and 1.75 Å for Pt, respectively^{34–36}, the calculated vdW gaps are 0.10–0.14 nm for the different metals used.

Additionally, the vdW gap can be directly determined from the cross-section TEM image (Fig. 1d) using the following equation:

$$T_{vdw} = d_{Mo-Au} - r_{Au} - r_S - d_{Mo-S} \quad (12)$$

where T_{vdw} is the vdW gap thickness, d_{Mo-Au} is the measured vdW distance between the Au surface plane and the molybdenum surface plane (about 0.53 nm as measured from Fig. 1d), r_{Au} is the Au atomic radius (0.135 nm), r_S is the sulfur atomic radius (0.088 nm) and d_{Mo-S} is the centre-to-centre distance between the molybdenum surface plane and the sulfur surface plane (0.162 nm).

The experimentally determined T_{vdw} is about 0.15 nm, consistent with theoretical expectations ($T_{theory} \approx 0.10$ – 0.14 nm). Such a thin tunnelling gap will result in a series resistance of around 10^{-10} Ω cm² to 10^{-8} Ω cm² according to equation (10), which is several orders of magnitude smaller than the typical MoS₂ contact resistance³⁷ (about 10^{-5} Ω cm² to 10^{-3} Ω cm²), and therefore can be largely neglected.

We note that the T_{vdw} value (0.15 nm) determined for the interface between the transferred metal and the semiconductor is comparable to values for other typical vdW interfaces. For example, the calculated the T_{vdw} gap between adjacent layers of graphene and MoS₂ is 0.20 nm and 0.15 nm, respectively, using equation (12) and the experimentally measured layer-to-layer distance (0.34 nm for graphite and 0.65 nm for molybdenite)^{38,39}.

Analysis of Ag–MoS₂–Pt MSM photodiode. To fabricate an Ag–MoS₂–Pt MSM photodiode, asymmetric electrode pairs consisting of Pt and Ag are first deposited on a sacrificial wafer and then released and physically laminated onto MoS₂ using the transfer method described above. To evaluate the photocurrent generation efficiency, we extract the photoresponsivity

$$R = I_{sc}/P_{laser} \quad (13)$$

where I_{sc} is the short-circuit current and P_{laser} is the input laser power. The measured R values are 7.2 mA W^{−1} for the monolayer and 16.6 mA W^{−1} for the seven-layer MoS₂ devices. With R determined, we can further extract the external quantum efficiency

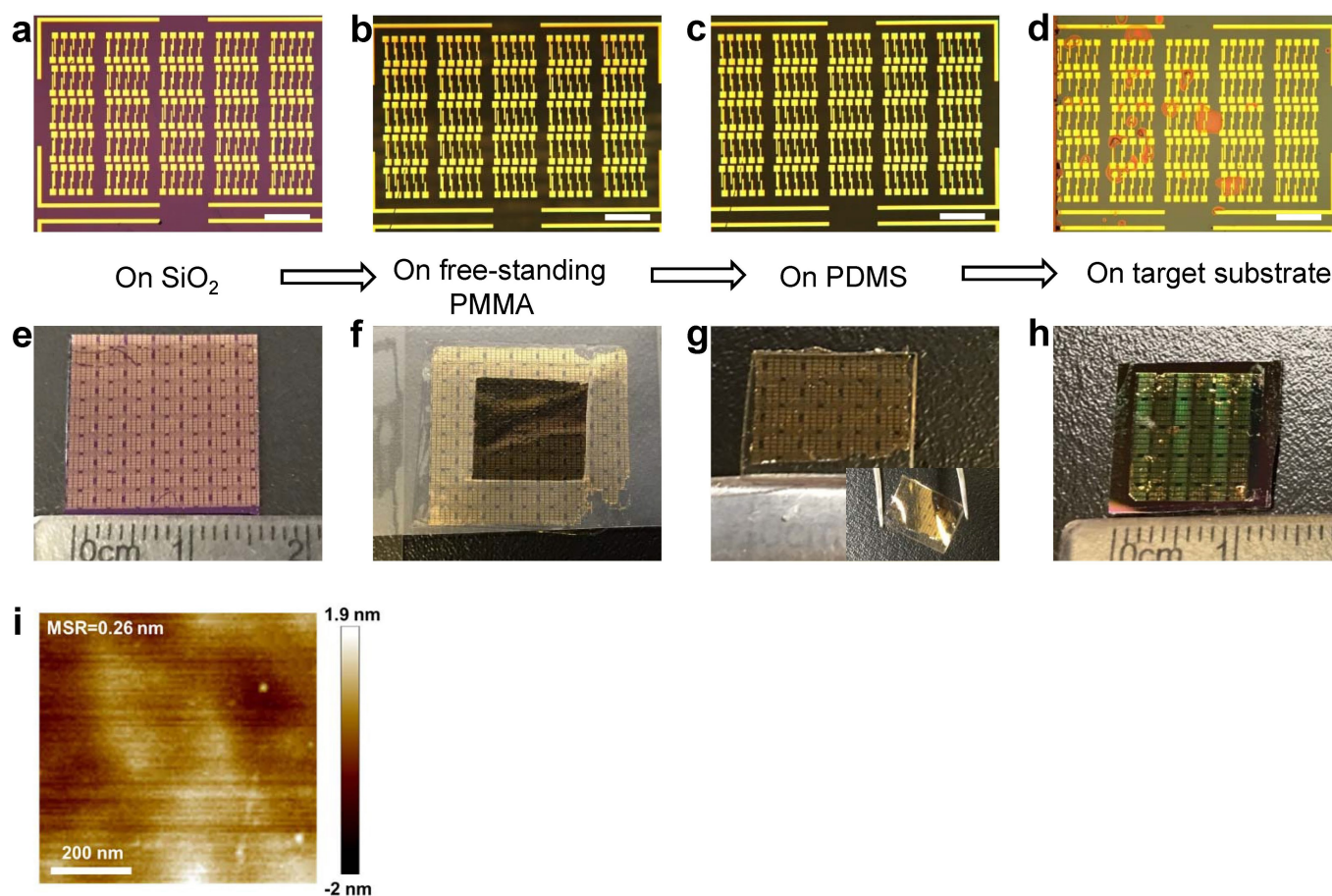
$$EQE = Rhc/e\lambda \quad (14)$$

where h , c , e and λ are Planck's constant, the speed of light, the electron charge and the laser wavelength, respectively. As the device produces both large I_{sc} and V_{oc} , the electrical power P_{el} can also be extracted. As shown in the dashed rectangular

area in Fig. 4d, a maximum electrical output power of 0.3 nW (3.7 nW) is obtained at $V_m = 0.54$ V (0.5 V) for the monolayer (multilayer) device. For the fill factor (FF), defined as the ratio of maximum obtainable power to the product of V_{oc} and short-circuit current I_{sc} , a value of $FF = P_{el,max}/(V_{oc}I_{sc}) \approx 0.26$ (0.47) is obtained for the monolayer (multi-layer) device. We can now also give an estimate of the power conversion efficiency, which is the percentage of the incident light energy that is converted into electrical energy, $\eta = P_{el,max}/P_{laser}$, where $\eta = 0.2\%$ and 0.61% for the monolayer and seven-layer devices, respectively.

Data availability. The data that support the findings of the current study are available from the corresponding authors upon reasonable request.

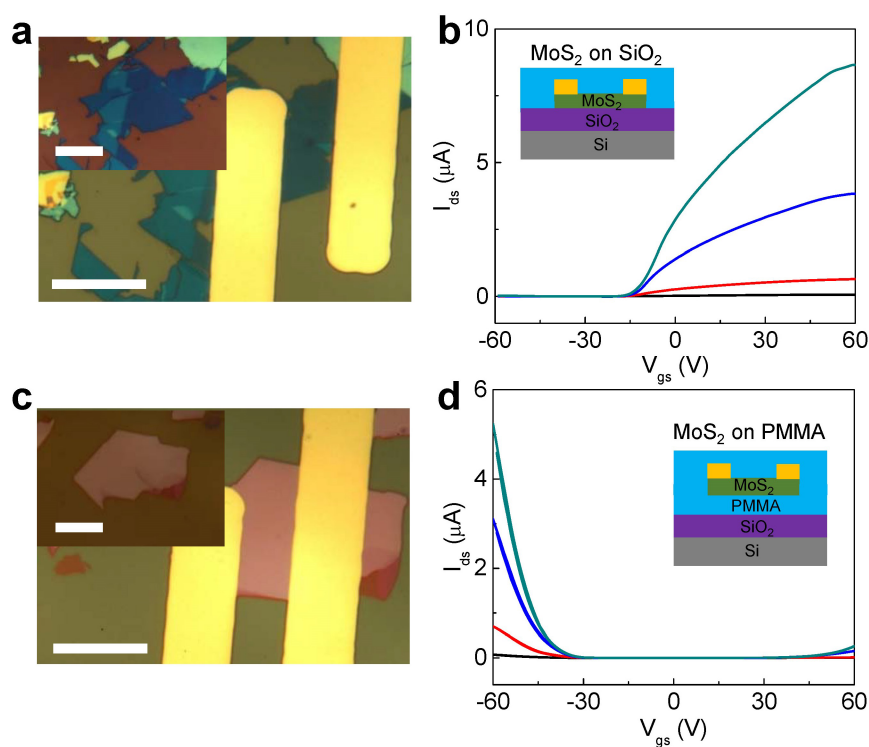
31. Yang, H. et al. Graphene barristor, a triode device with a gate-controlled Schottky barrier. *Science* **336**, 1140–1143 (2012).
32. Allain, A., Kang, J., Banerjee, K. & Kis, A. Electrical contacts to two-dimensional semiconductors. *Nat. Mater.* **14**, 1195–1205 (2015).
33. Ranuárez, J. C., Deen, M. J. & Chen, C.-H. A review of gate tunneling current in MOS devices. *Microelectron. Reliab.* **46**, 1939–1956 (2006).
34. Slater, J. C. Atomic radii in crystals. *J. Chem. Phys.* **41**, 3199–3204 (1964).
35. Bondi, A. Van der Waals volumes and radii. *J. Phys. Chem.* **68**, 441–451 (1964).
36. Mantina, M., Chamberlin, A. C., Valero, R., Cramer, C. J. & Truhlar, D. G. Consistent van der Waals radii for the whole main group. *J. Phys. Chem. A* **113**, 5806–5812 (2009).
37. Lee, S., Tang, A., Aloni, S. & Wong, H.-S. P. Statistical study on the Schottky barrier reduction of tunneling contacts to CVD synthesized MoS₂. *Nano Lett.* **16**, 276–281 (2016).
38. Delhaes, P. Graphite and Precursors Vol. 1 (CRC, London, 2000).
39. Radisavljevic, B., Radenovic, A., Brivio, J., Giacometti, V. & Kis, A. Single-layer MoS₂ transistors. *Nat. Nanotech.* **6**, 147–150 (2011).
40. Li, X. et al. Performance potential and limit of MoS₂ transistors. *Adv. Mater.* **27**, 1547–1552 (2015).
41. Yang, L. et al. High-performance MoS₂ field-effect transistors enabled by chloride doping: record low contact resistance (0.5 k Ω · μ m) and record high drain current (460 μ A/ μ m). In *Symp. VLSI Technology and Circuits (VLSI 2014)* 192–193 (2014).
42. Kappera, R. et al. Phase-engineered low-resistance contacts for ultrathin MoS₂ transistors. *Nat. Mater.* **13**, 1128–1134 (2014).
43. Liu, Y. et al. Pushing the performance limit of sub-100 nm molybdenum disulfide transistors. *Nano Lett.* **16**, 6337–6342 (2016).
44. Chuang, S. et al. MoS₂ p-type transistors and diodes enabled by high work function MoO_x contacts. *Nano Lett.* **14**, 1337–1342 (2014).
45. Liu, X. et al. P-type polar transition of chemically doped multilayer MoS₂ transistor. *Adv. Mater.* **28**, 2345–2351 (2016).
46. Laskar, M. R. et al. P-type doping of MoS₂ thin films using Nb. *Appl. Phys. Lett.* **104**, 092104 (2014).
47. Fontana, M. et al. Electron–hole transport and photovoltaic effect in gated MoS₂ Schottky junctions. *Sci. Rep.* **3**, 1634 (2013); corrigendum **5**, 12589 (2015).
48. Lee, C.-H. et al. Atomically thin p–n junctions with van der Waals heterointerfaces. *Nat. Nanotech.* **9**, 676–681 (2014).
49. Cheng, R. et al. Electroluminescence and photocurrent generation from atomically sharp WSe₂/MoS₂ heterojunction p–n diodes. *Nano Lett.* **14**, 5590–5597 (2014).
50. Furchi, M. M., Pospischil, A., Libisch, F., Burgdörfer, J. & Mueller, T. Photovoltaic effect in an electrically tunable van der Waals heterojunction. *Nano Lett.* **14**, 4785–4791 (2014).
51. Deng, Y. et al. Black phosphorus–monolayer MoS₂ van der Waals heterojunction p–n diode. *ACS Nano* **8**, 8292–8299 (2014).
52. Li, D. et al. Two-dimensional non-volatile programmable p–n junctions. *Nat. Nanotech.* **12**, 901–906 (2017).
53. Yu, W. J. et al. Highly efficient gate-tunable photocurrent generation in vertical heterostructures of layered materials. *Nat. Nanotech.* **8**, 952–958 (2013).



Extended Data Fig. 1 | Optical images, photographs and characterization of the transfer process of the metal electrodes.

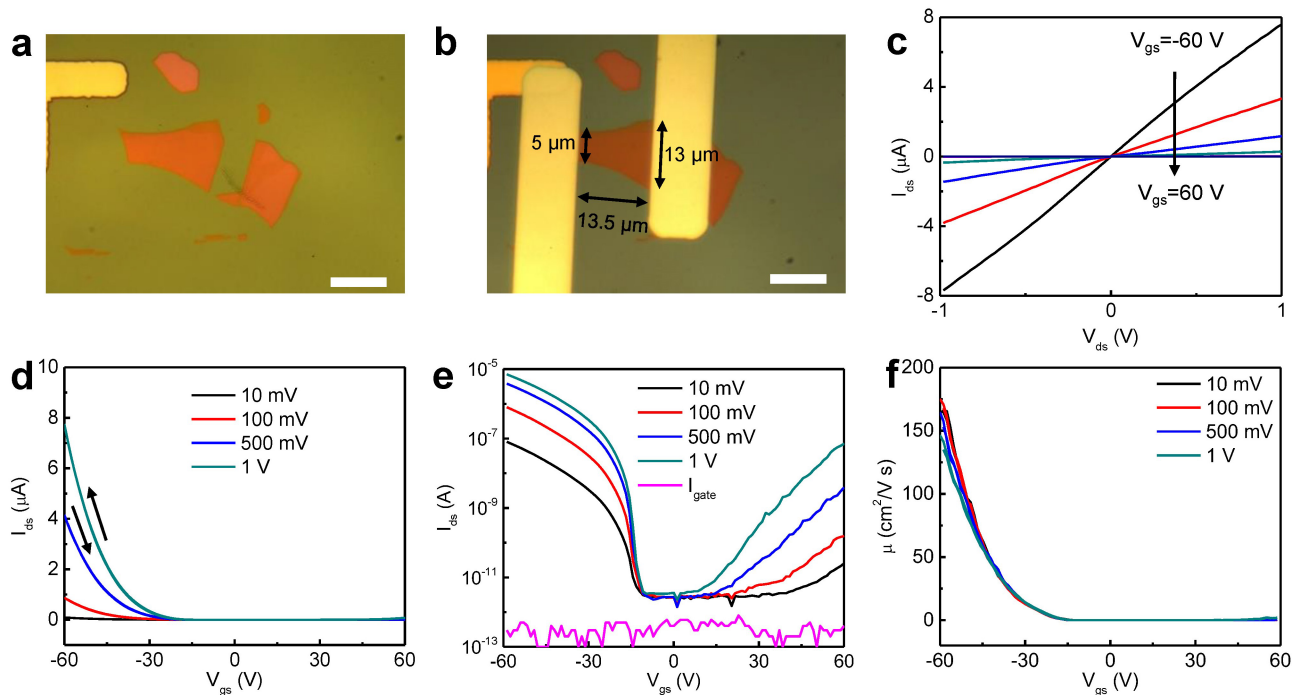
a–d, Optical images of Au electrodes deposited on a SiO_2 substrate (**a**), physically released using 1- μm -thick PMMA (**b**), attached to a PDMS (with PMMA) substrate (**c**) and transferred onto the target substrate (**d**). Scale bars, 200 μm in **a–d**. **e–h**, The corresponding photographs of Au

electrodes deposited on a SiO_2 substrate (**e**), physically released using 1- μm -thick PMMA (**f**), attached on a PDMS (with PMMA) substrate (**g**), and transferred onto the target substrate (**h**). **i**, Atomic force microscopy image of the bottom side of the transferred electrodes, with a root-mean-square surface roughness of 0.26 nm.



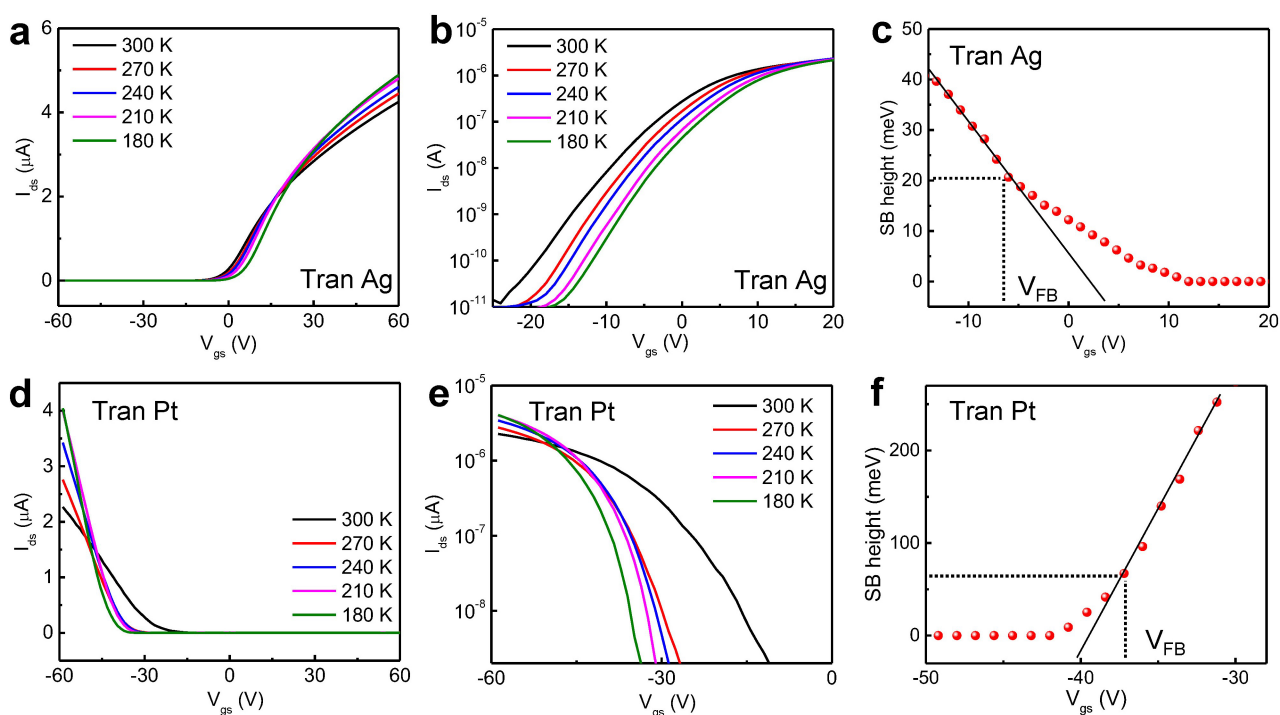
Extended Data Fig. 2 | Substrate doping effect on MoS₂. **a**, Optical image of a seven-layer MoS₂ flake on a SiO₂ substrate contacted with transferred Pt electrodes. Inset, the optical image of MoS₂ on SiO₂ before the metal contact. Scale bar, 20 μm. **b**, I_{ds} - V_{gs} transfer curve of MoS₂ transistor on a SiO₂ substrate under various bias voltages of 10 mV (black), 100 mV (red), 500 mV (blue) and 1 V (cyan), demonstrating n-type behaviour, suggesting the involvement of defect states within the SiO₂-MoS₂ interface. **c**, Optical image of a MoS₂ flake approximately 15 layers thick

on a PMMA substrate, contacted with transferred Pt electrodes. Inset, the optical image of MoS₂ on PMMA before the metal contact. Scale bar, 20 μm. **d**, I_{ds} - V_{gs} transfer curve of MoS₂ transistor encapsulated in PMMA under various bias voltages of 10 mV (black), 100 mV (red), 500 mV (blue) and 1 V (cyan), demonstrating p-type behaviour, suggesting that the use of a PMMA substrate is essential for preventing substrate pinning effects and retaining the intrinsic properties of MoS₂ flakes. All measurements were conducted at room temperature in probe stations.



Extended Data Fig. 3 | Highest-hole-mobility device using transferred Pt as the contact electrodes. **a**, Optical image of a MoS₂ flake on a PMMA/SiO₂ substrate. **b**, Optical image of the MoS₂ flake after being contacted by transferred Pt electrodes. The channel length is 13.5 μm and the effective channel width is 8.37 μm. Scale bar in **a**, **b**, 10 μm. **c**, I_{ds} - V_{ds} output curve of the MoS₂ transistor under various gate voltages from -60 V to 60 V. **d**, **e**, Linear (**d**) and semi-logarithmic (**e**) plot of the I_{ds} - V_{gs} transfer curve of the MoS₂ transistor under various bias voltages: 10 mV (black), 100 mV (red), 500 mV (blue) and 1 V (cyan). The purple line is the gate leakage current (I_g), which is an order of magnitude smaller (limited by equipment) than I_{ds} and will not affect the

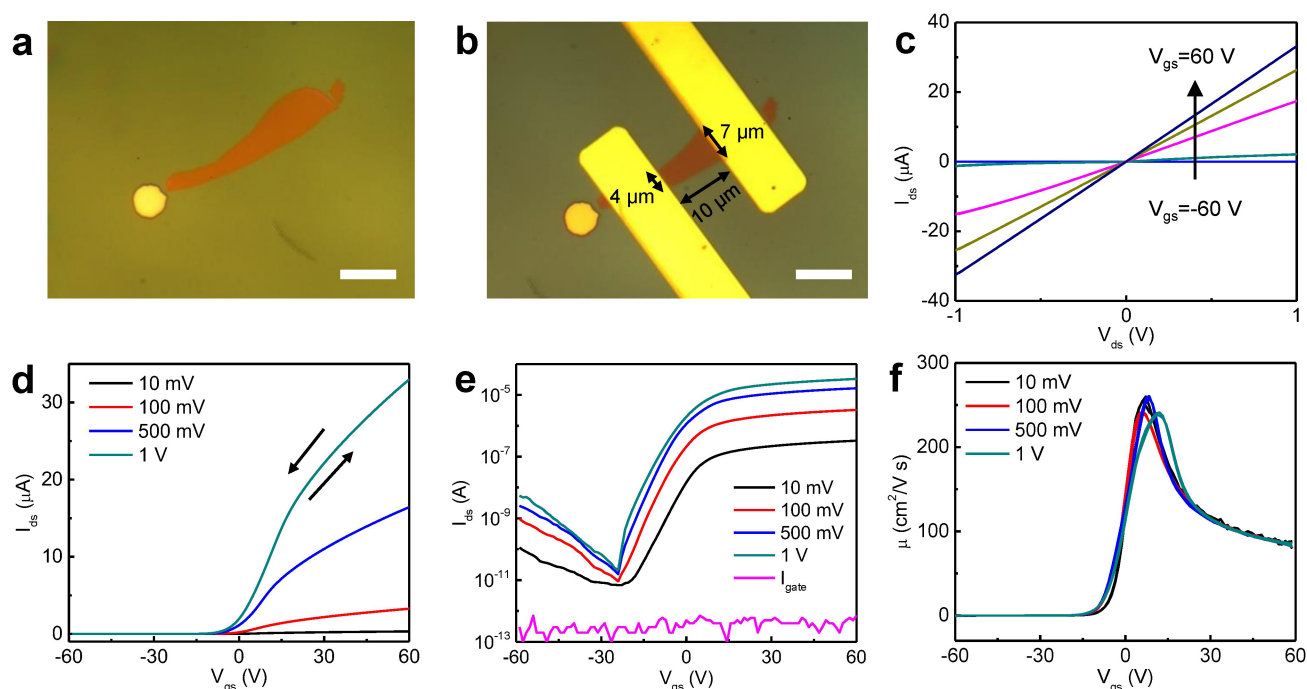
overall carrier transport. Under large gate voltage, the channel majority carrier is inverted to electrons and the carrier concentration is increased exponentially, greatly reducing the electron Schottky barrier width. As a result, the electrons can tunnel through the thin Schottky barrier from the source side, which accounts for the observed ambipolar behaviour. **f**, The extracted two-terminal field-effect hole mobility using various bias voltages: 10 mV (black), 100 mV (red), 500 mV (blue), 1 V (cyan). The width/length ratio is 0.62. The gate dielectric is composed of 300-nm-thick SiO₂ and 170-nm-thick PMMA and is calculated to be 6.2 nF cm⁻². The highest extracted hole mobility is 175 cm² V⁻¹ s⁻¹. All measurements were conducted at room temperature in probe stations.



Extended Data Fig. 4 | Flat-band Schottky barrier extraction.

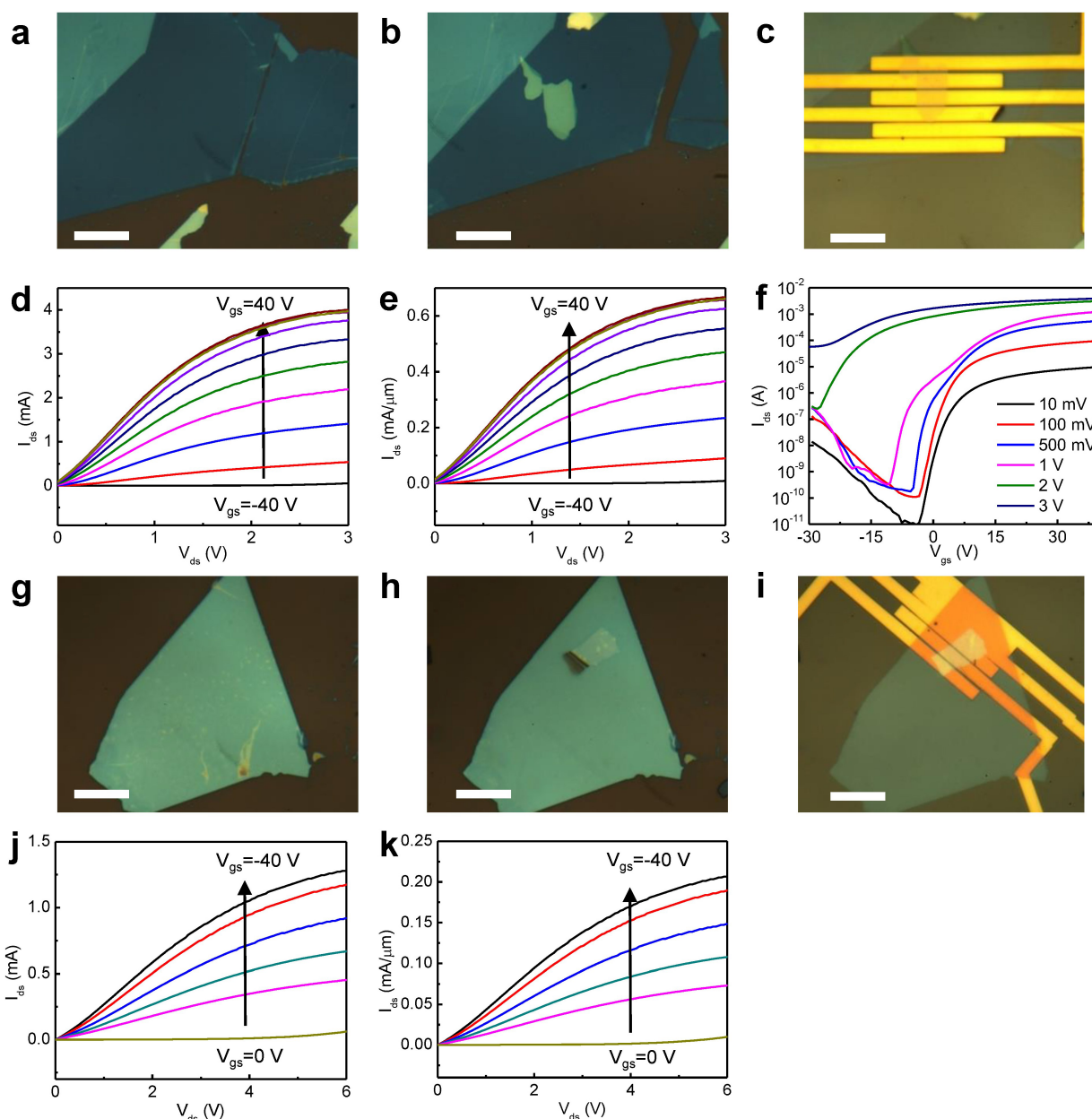
a, b, I_{ds} - V_{gs} transfer curves of a MoS₂ transistor using transferred Ag electrodes under various temperatures, with the bias voltage fixed at 100 mV. **c,** The extracted n-type Schottky barrier height at various gate voltages, where the flat-band electron Schottky barrier is measured to be 20 mV. The flat-band voltage and corresponding Schottky barrier are

shown by the dashed lines. **d, e,** I_{ds} - V_{gs} transfer curves of a MoS₂ transistor using transferred Pt electrodes under various temperatures, with the bias voltage fixed at 100 mV. **f,** The extracted p-type Schottky barrier height at various gate voltages, where the flat-band hole Schottky barrier is measured to be 67 mV. The flat-band voltage and corresponding Schottky barrier are shown by the dashed lines. Tran, transferred.



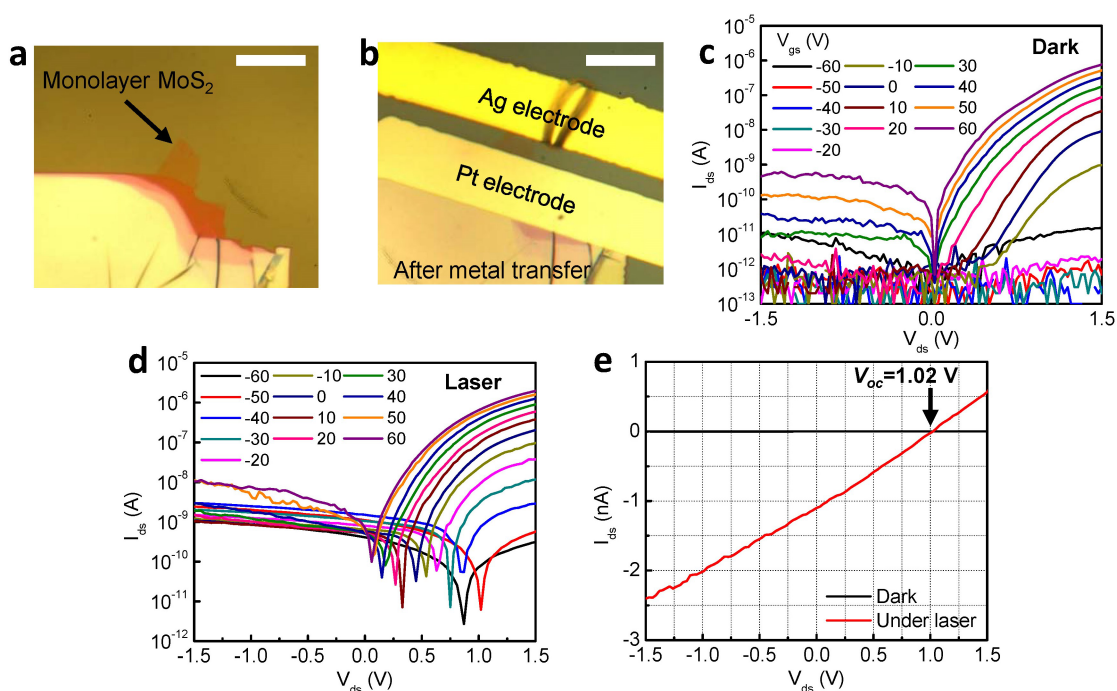
Extended Data Fig. 5 | Highest-electron-mobility device using transferred Ag as the contact electrodes. **a**, Optical image of a MoS₂ flake on a PMMA/SiO₂ substrate. **b**, Optical image of the MoS₂ flake after being contacted by transferred Ag electrodes. The channel length here is 10 μm and the effective channel width is 5.36 μm. Scale bar in **a**, **b**, 10 μm. **c**, I_{ds} - V_{ds} output curve of the MoS₂ transistor under various gate voltages from -60 V to 60 V. **d**, **e**, Linear (**d**) and semi-logarithmic (**e**) plot of I_{ds} - V_{gs} transfer curve of the MoS₂ transistor under various bias voltages: 10 mV (black), 100 mV (red), 500 mV (blue) and 1 V (cyan). The purple line is the gate leakage current (I_g), which is an order of magnitude smaller than I_{ds} (limited by equipment) and will not affect the overall carrier transport.

Under large gate voltage, the channel majority carrier is inverted to holes and the carrier concentration is increased exponentially, greatly reducing the hole Schottky barrier width. As a result, the holes can tunnel through the thin Schottky barrier from the drain side, which accounts for the observed ambipolar behaviour. **f**, The extracted two-terminal field-effect electron mobility using various bias voltages: 10 mV (black), 100 mV (red), 500 mV (blue) and 1 V (cyan). The width/length ratio is 0.54. The gate dielectric is composed of 300-nm-thick SiO₂ and 170-nm-thick PMMA and is calculated to be 6.2 nF cm⁻². The highest extracted electron mobility is 260 cm² V⁻¹ s⁻¹. All measurements are conducted at room temperature in probe stations.



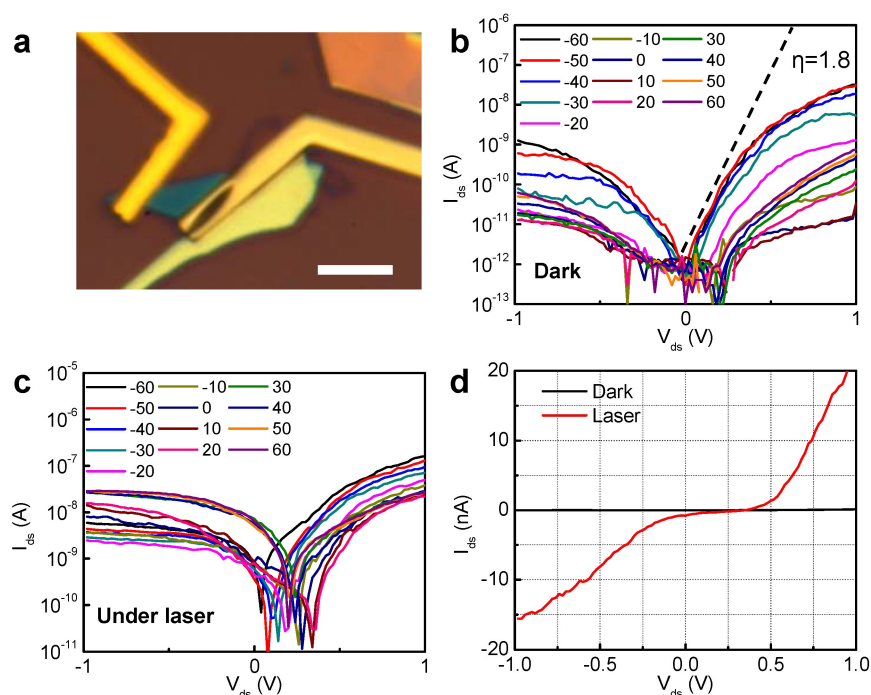
Extended Data Fig. 6 | Highest n-type current density using transferred Ag and p-type current density using transferred Pt as the contact electrodes. **a–c**, Optical image of initial thin BN flake (**a**), after MoS₂ has been dry-transferred onto BN using an alignment transfer technique (**b**), and the final device with transferred Ag electrodes (**c**). The channel length is about 160 nm and the channel width is about 6 μm . The gate dielectric is composed of approximately 5-nm-thick BN flake and 90-nm-thick SiO₂ (rather than the 300-nm-thick SiO₂ and 170-nm-thick PMMA dielectric used previously) for larger gate capacitance and stronger gate coupling to ensure the highest driving current. **d, e**, I_{ds} - V_{ds} output curves of the fabricated MoS₂ transistor under various gate voltages from -40 V to 40 V. The highest current density is measured to be 0.66 $\text{mA } \mu\text{m}^{-1}$.

f, I_{ds} - V_{gs} transfer curve of the fabricated MoS₂ transistor under various bias voltages. With increasing bias voltage, the OFF current of the device increases owing to the short channel effect. **g–i**, Optical image of initial thin BN flake (**g**), after MoS₂ has been dry-transferred onto BN using an alignment transfer technique (**h**), and the final device with transferred Pt electrodes (**i**). The channel length is ~ 140 nm, the channel width is about 6 μm and the gate dielectric is composed of approximately 10-nm-thick BN flake and 90-nm-thick SiO₂. **j, k**, I_{ds} - V_{ds} output curve of the fabricated MoS₂ transistor under various gate voltages from 0 V to -40 V. The highest current density is measured to be 0.21 $\text{mA } \mu\text{m}^{-1}$. Scale bar in **a–c** and **g–i**, 10 μm . All measurements were conducted at room temperature in probe stations.



Extended Data Fig. 7 | Photoresponse of a monolayer MoS₂ device with transferred Ag and Pt asymmetric electrodes. **a**, Optical image of monolayer MoS₂ mechanically exfoliated on a 170 nm PMMA/300 nm SiO₂ substrate. **b**, Optical image of the device after Ag and Pt asymmetric electrodes are transferred on top of monolayer MoS₂. Scale bar in **a**, **b**, 10 μ m. **c**, Semi-logarithmic plot of I_{ds} - V_{ds} output curve under various

gate voltages (-60 V to 60 V, 10 V step) under dark conditions. The Pt is biased and the Ag is grounded. **d**, Semi-logarithmic plot of I_{ds} - V_{ds} output curve under various gate voltages (-60 V to 60 V, 10 V step) under laser illumination. **e**, The I_{ds} - V_{ds} output curve under dark and laser illumination, under gate -50 V. The highest open-circuit voltage of 1.02 V is observed in monolayer devices.



Extended Data Fig. 8 | Photoresponse of a multilayer MoS₂ device with deposited Ag and Pt asymmetric electrodes. **a**, Optical image of the device. Scale bar, 5 μm . **b**, Semi-logarithmic plot of $I_{\text{ds}}-V_{\text{ds}}$ output curve under various gate voltages (-60 V to 60 V, 10 V step) under dark conditions. The Pt is biased and the Ag is grounded. **c**, Semi-logarithmic

plot of $I_{\text{ds}}-V_{\text{ds}}$ output curve under various gate voltage (-60 V to 60 V, 10 V step) under laser illumination. **d**, $I_{\text{ds}}-V_{\text{ds}}$ output curve under dark conditions and laser illumination, under a gate voltage of 10 V. The highest open-circuit voltage of about 0.3 V is observed.

Extended Data Table 1 | Electrical performance of MoS₂ devices

Method	* L_{ch} (nm)	I_{on} (mA/ μ m)	*Extrinsic mobility (cm ² /V s), RT	Reference
		n-type		
Transferred Ag contact	160	0.66	260 (electrons)	This work
Sc contact	5000	0.24	184 (electrons)	Ref. 24
Contact Cl doping	100	0.33	50 (electrons)	Ref. 40
Contact DEC doping	100	0.46	55 (electrons)	Ref. 41
Contact phase engineering	1200	0.085	46 (electrons)	Ref. 42
Graphene contact	~3000	0.01	70 (electrons)	Ref. 26
Metal/graphene vdW contact [#]	80	0.83	51 (electrons)	Ref. 43
		p-type		
Transferred Pt contact	140	0.21	175 (holes)	This work
Pd/MO _x contact	7000	0.00014	<1 (holes)	Ref. 44
AuCl ₃ doping	~1500	0.02	68 (holes)	Ref. 45
Nb doping	NA	NA	8.5 (holes, Hall)	Ref. 46

*The channel lengths (L_{ch}) here are for the device with highest current density.

+The mobility is extracted through measurements with a field effect transistor (FET mobility), except for the Nb doping device, which uses Hall measurement (Hall mobility).

#The only higher I_{on} is achieved in an ultra-short n-channel MoS₂ device with optimized metal/graphene hybrid vdW contact.

n-type results are from refs. ^{24,26,40–43}; p-type results from refs. ^{44–46}.

Extended Data Table 2 | Photovoltaic effect in 2D semiconductor-based diodes

Method	2D thickness	V_{oc} (V)	Reference
Transferred Ag-MoS ₂ -Ag diode	1 layer	1.02	This work
Transferred Ag-MoS ₂ -Ag diode	7 layer	0.76	This work
Evaporated Pd-MoS ₂ -Au diode	Multilayer	0.1	Ref. 47
Dual-gate WSe ₂ p-n diode	1 layer	0.86	Ref. 29
Dual-gate WSe ₂ p-n diode	1 layer	0.7	Ref. 30
MoS ₂ -WSe ₂ vertical p-n diode	9 layer-10 layer	0.5	Ref. 48
MoS ₂ -WSe ₂ planar p-n diode	Few layer-1 layer	0.3	Ref. 49
MoS ₂ -WSe ₂ planar p-n diode	1 layer-1 layer	0.59	Ref. 50
MoS ₂ -BP planar p-n diode	1 layer-few layer	0.27	Ref. 51
WSe ₂ -BN-graphene diode	3 layer-20 layer-3 layer	0.82	Ref. 52
Graphene-MoS ₂ -graphene vertical diode	1 layer-multilayer-1 layer	0.3	Ref. 53

Data are from refs.^{29,30,47–53}.

The effect of hydration number on the interfacial transport of sodium ions

Jinbo Peng^{1,9,10}, Duanyun Cao^{1,10}, Zhili He^{2,10}, Jing Guo¹, Prokop Hapala³, Runze Ma¹, Bowei Cheng¹, Ji Chen⁴, Wen Jun Xie², Xin-Zheng Li^{5,6}, Pavel Jelínek^{3,7}, Li-Mei Xu^{1,6*}, Yi Qin Gao^{2*}, En-Ge Wang^{1,6,8*} & Ying Jiang^{1,6,8*}

Ion hydration and transport at interfaces are relevant to a wide range of applied fields and natural processes^{1–5}. Interfacial effects are particularly profound in confined geometries such as nanometre-sized channels^{6–8}, where the mechanisms of ion transport in bulk solutions may not apply^{9,10}. To correlate atomic structure with the transport properties of hydrated ions, both the interfacial inhomogeneity and the complex competing interactions among ions, water and surfaces require detailed molecular-level characterization. Here we constructed individual sodium ion (Na⁺) hydrates on a NaCl(001) surface by progressively attaching single water molecules (one to five) to the Na⁺ ion using a combined scanning tunnelling microscopy and noncontact atomic force microscopy system. We found that the Na⁺ ion hydrated with three water molecules diffuses orders of magnitude more quickly than other ion hydrates. Ab initio calculations revealed that such high ion mobility arises from the existence of a metastable state, in which the three water molecules around the Na⁺ ion can rotate collectively with a rather small energy barrier. This scenario would apply even at room temperature according to our classical molecular dynamics simulations. Our work suggests that anomalously high diffusion rates for specific hydration numbers of ions are generally determined by the degree of symmetry match between the hydrates and the surface lattice.

Determination of molecular-level details of hydration processes remains a great challenge, both experimentally and theoretically. Various spectroscopic techniques have been used to identify the structure and dynamics of solvated ions or molecules through vibrational fingerprints^{11–13}. However, all of these techniques suffer from poor spatial resolution and the difficulty of spectral assignment. Molecular simulations have also become powerful tools with which to investigate atomic-scale hydration properties^{2,14,15}, but the reliability of the results depends critically on many tunable factors¹⁴. Recently, scanning probe microscopy has provided an opportunity to probe interfacial water at the single-molecule or even submolecular level^{16–22}, but the application to ion hydration systems is not straightforward owing to the lack of controlled methods of preparing individual ion hydrates on the substrates and the high flexibility of their structures^{23,24}. Using an ultrahigh-resolution scanning tunnelling microscopy (STM) and qPlus²⁵ noncontact atomic force microscopy (AFM) combined system, here we studied the hydrated Na⁺ ion, an alkali metal ion abundant in natural water and biological solutions.

The Na⁺ hydrates (Na⁺·*n*D₂O, *n* = 1–5) were assembled in a controlled manner by progressively attaching single D₂O molecules to Na⁺ ions, which were extracted from the NaCl surface with a chlorine (Cl)[–]-terminated tip (for detailed procedures, see Methods and Supplementary Fig. 1). We found that the barrier for extracting the Na⁺ from NaCl was greatly suppressed with the assistance of water²⁶.

Figure 1a–e shows the atomic models, STM/AFM images (acquired with a CO tip²⁷) and AFM simulations of Na⁺·*n*D₂O clusters (*n* = 1–5). From the atomically resolved STM images, it can be clearly seen that Na⁺·D₂O, Na⁺·2D₂O, and Na⁺·3D₂O adsorb at the bridge sites, while Na⁺·4D₂O and Na⁺·5D₂O adsorb on top of the Cl[–] sites. We found that the maximum number of water molecules in the first hydration shell is five (see Supplementary Fig. 2). Further addition of water to Na⁺·5D₂O results in formation of the second and higher hydration shells.

The AFM images of the ion hydrates were acquired within the weak-disturbance region where the high-order electrostatic force is dominant²¹, providing higher resolution and finer details than STM. The charge state of Na in the hydrates can be verified by comparing the AFM images and simulations (Supplementary Fig. 3). In the AFM images, the Na⁺ ion appears as a dark depression, mainly arising from the electrostatic attraction between the CO-tip apex and the Na⁺ ion. By contrast, the water molecule was imaged as a bright feature (white arrow in Fig. 1a) surrounded by a dark ring (white dashed curve in Fig. 1a), which are ascribed to the negatively charged O atom and the positively charged D atom, respectively²¹. The ‘standing’ water (that is, the molecular plane of the water molecule is perpendicular to the surface, in contrast to the flat-lying water molecules) of Na⁺·3D₂O (see the white arrow in Fig. 1c) shows a prominent protrusion caused by the Pauli repulsion force. The fuzzy feature in the AFM image of Na⁺·D₂O (see the blue arrow in Fig. 1a) may result from the flipping of water over Na⁺ in the presence of the tip (for more experimental evidence, see Supplementary Fig. 4). The AFM simulations based on a molecular mechanics model using a quadrupole tip (Methods) nicely reproduce all the experimental images. The comparison between the submolecular-resolution AFM image and simulation is important in determining the structure of ion hydrates (one example is shown in Supplementary Fig. 5).

Next we explored the transport of those hydrates. To activate their diffusion at low temperature (5 K), we used the inelastic electron tunnelling technique by injecting ‘hot’ (that is, with larger energy than those at the experimental temperature) electrons/holes into the Au substrate, which transport along the surface and transfer their energy to the hydrates^{28,29} (Fig. 2a). Figure 2b plots the diffusion probability of Na⁺·3D₂O and Na⁺·3H₂O as a function of the bias voltage. It clearly shows a fast increase around ±150 meV (±170 mV), which coincides with the bending mode of D₂O (H₂O). Therefore, the diffusion of Na⁺ hydrates must have been induced by the vibrational excitation in the one-electron process (Fig. 2b, inset). The diffusion direction is almost random when using a CO tip. However, the Na⁺ hydrates tend to diffuse towards the Cl[–] tip, owing to the electrostatic attraction between the Na⁺ and the Cl[–] at the tip apex (see Supplementary Fig. 6 for experimental evidence and theoretical support). This provides a convenient way to study the diffusion of hydrates.

¹International Center for Quantum Materials, School of Physics, Peking University, Beijing, China. ²Institute of Theoretical and Computational Chemistry, College of Chemistry and Molecular Engineering, Peking University, Beijing, China. ³Institute of Physics, Czech Academy of Sciences, Prague, Czech Republic. ⁴Department of Physics and Astronomy, London Centre for Nanotechnology, Thomas Young Centre, University College London, London, UK. ⁵State Key Laboratory for Mesoscopic Physics and School of Physics, Peking University, Beijing, China. ⁶Collaborative Innovation Center of Quantum Matter, Beijing, China. ⁷Regional Centre of Advanced Technologies and Materials, Palacký University, Olomouc, Czech Republic. ⁸CAS Center for Excellence in Topological Quantum Computation, University of Chinese Academy of Sciences, Beijing, China. ⁹Present address: Institute of Experimental and Applied Physics, University of Regensburg, Regensburg, Germany. ¹⁰These authors contributed equally: Jinbo Peng, Duanyun Cao, Zhili He. *e-mail: limei.xu@pku.edu.cn; gaoyq@pku.edu.cn; egwang@pku.edu.cn; yjiang@pku.edu.cn

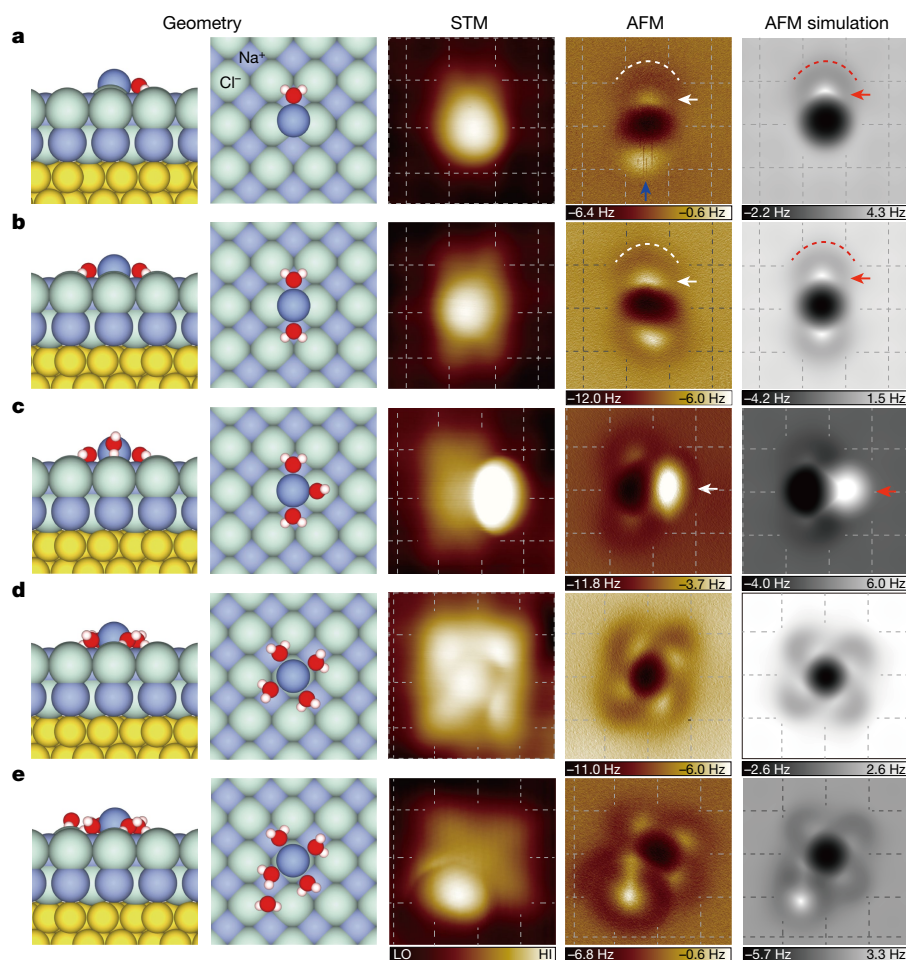


Fig. 1 | Geometries and high-resolution STM/AFM images of Na^+ hydrates. **a–e**, The atomic models (the first column shows the side view; the second column shows the top view), STM and AFM images (acquired with a CO tip) and AFM simulations of $\text{Na}^+ \cdot n\text{D}_2\text{O}$ clusters ($n = 1–5$), respectively. H, O, Cl, Na and Au atoms are denoted as white, red, green, purple and yellow spheres, respectively. Square lattices of the NaCl(001) surface arising from the Cl^- are depicted in the STM/AFM/simulation images by dashed grids. The white (red) arrows in **a** and **b** denote bright protrusions, and the white (red) dashed curves highlight the crooked depressions in the AFM images (AFM simulations). The blue arrow in **a** denotes the fuzzy feature arising from the flipping of the water molecule. The white (red) arrow in **c** denotes the standing water in the AFM image (AFM simulation). The set points of the STM images (**a–e**) are $V = 100$ mV

and $I = 20$ pA; $V = 150$ mV and $I = 30$ pA; $V = 100$ mV and $I = 30$ pA; $V = 100$ mV and $I = 50$ pA; and $V = 100$ mV and $I = 15$ pA, respectively. The tip heights of experimental (simulated) AFM images (**a–e**) are 130 pm (7.90 Å), 80 pm (8.10 Å), 100 pm (7.95 Å), 100 pm (8.10 Å) and 100 pm (7.99 Å), respectively. The tip height of experimental AFM images is referenced to the STM set point on the NaCl surface (100 mV, 50 pA). The tip height in simulations is defined as the vertical distance between the apex atom of the metal tip and the Na^+ ion in Na^+ hydrates. All the AFM oscillation amplitudes of experimental and simulated images are 100 pm. All the AFM simulations were done with a quadrupole (d_z^2) tip ($k = 0.75 \text{ N m}^{-1}$, $Q = -0.2e$, where Q is the magnitude of quadrupole charge at the tip apex and e is the elementary charge). The images are $1.5 \text{ nm} \times 1.5 \text{ nm}$.

To compare the mobility of different Na^+ hydrates, we adopted the following procedures: first, the Cl^- tip was positioned away from the hydrates at a certain lateral distance ($d \times$ the lattice constant of NaCl(001), which is 0.39 nm); second, the bias voltage was ramped slowly while the tip height was kept constant; and third, the tunnelling current experienced a sudden increase at a certain effective bias voltage (V_{eff}), signalling that the hydrates had reached the tip (Fig. 2c). We found that the as-determined V_{eff} increases as d increases (Fig. 2d). The behaviour of V_{eff} shows a negligible difference between the positive and negative biases, again revealing the critical role of the vibrational excitation (Fig. 2d).

Although the V_{eff} does not simply represent the diffusion barrier and is subject to various experimental parameters (Supplementary Fig. 7), we can still use this quantity to compare the relative mobility of different hydrates in a qualitative way. Figure 2e plots the V_{eff} as a function of d for different hydrates. We notice that $\text{Na}^+ \cdot 3\text{D}_2\text{O}$ has a much smaller V_{eff} than other hydrates. The hydrates never reached the tip for $d > 2$ even when the bias increased to 700 mV, except for $\text{Na}^+ \cdot 3\text{D}_2\text{O}$. We found that the tip may induce large structural change for $\text{Na}^+ \cdot 4\text{D}_2\text{O}$ and $\text{Na}^+ \cdot 5\text{D}_2\text{O}$ at

$d = 2$ (Supplementary Fig. 8), leading to a considerable reduction of V_{eff} . Strikingly, the $\text{Na}^+ \cdot 3\text{D}_2\text{O}$ can still diffuse to the tip readily at $d = 7$, with a relatively small V_{eff} (about 400 mV) (Fig. 2d). This suggests that the $\text{Na}^+ \cdot 3\text{D}_2\text{O}$ may have an unusually small diffusion barrier.

To gain insights into the diffusion pathway of those hydrates, we performed ab initio density functional theory (DFT) calculations (Methods). Indeed, the $\text{Na}^+ \cdot 3\text{H}_2\text{O}$ has the lowest diffusion barrier (below 80 meV) and the potential energy landscape along the path from the Cl^- bridge to Cl^- atop is rather flat, while all the other hydrates have barriers well above 200 meV (Fig. 3a). The initial, transition and final states are depicted in Fig. 3b. The diffusion of $\text{Na}^+ \cdot n\text{H}_2\text{O}$ ($n = 1–3$) is accompanied with the rotation of water around the Na^+ , whereas $\text{Na}^+ \cdot n\text{H}_2\text{O}$ ($n = 4$ and 5) follows a translational mode only with local rearrangements of water. The translational diffusion barrier of $\text{Na}^+ \cdot 3\text{H}_2\text{O}$ is almost three times that of the rotational one, while for $\text{Na}^+ \cdot 2\text{H}_2\text{O}$ and $\text{Na}^+ \cdot \text{H}_2\text{O}$ the barriers of the two modes are nearly the same (Supplementary Fig. 9).

The small diffusion barrier of $\text{Na}^+ \cdot 3\text{H}_2\text{O}$ is closely related to the existence of a peculiar metastable state (T3 in Fig. 3b), where the Na^+

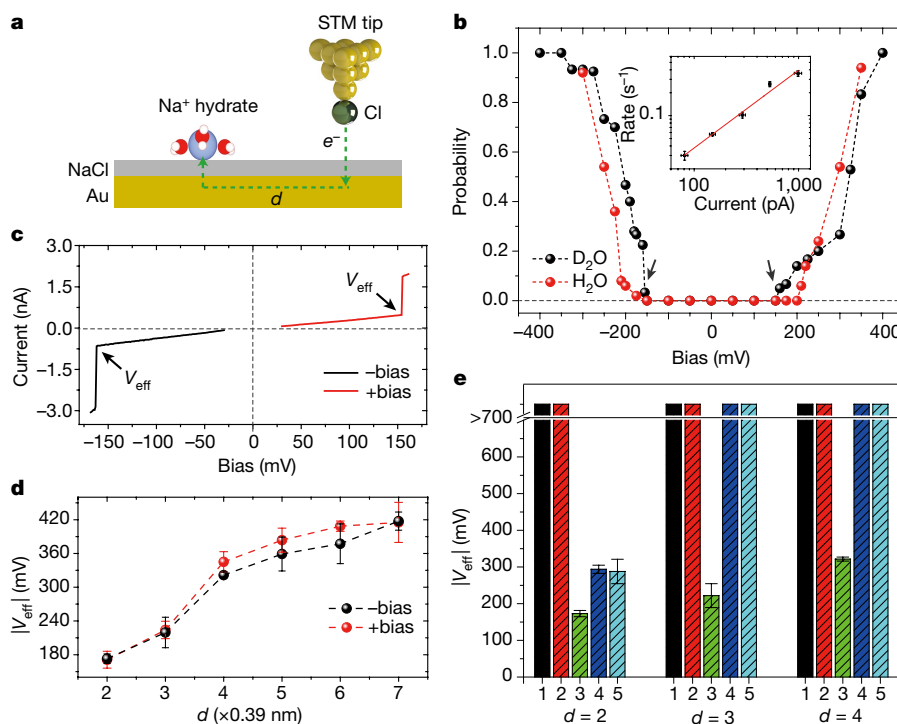


Fig. 2 | Tip-induced diffusion dynamics of Na⁺ hydrates. **a**, Schematic diagram of the Au-mediated inelastic electron excitation of the Na⁺ hydrates with the STM tip at a lateral distance of $d \times$ the lattice constant of NaCl(001), which is 0.39 nm. The flow direction of hot electrons is indicated by the green dotted arrow. **b**, Bias dependence of the diffusion probability of Na⁺·3D₂O and Na⁺·3H₂O with a CO tip at $d = 4$. The voltage pulse duration for each event is 1.2 s. The diffusion probability is a statistics from 50 events. The threshold bias for D₂O is indicated by two black arrows. The inset shows the current dependence of the diffusion rate of Na⁺·3D₂O with a CO tip at $d = 2$ under 170 mV. Error bars of

the current reflect the standard deviation from the set point. Error bars of the diffusion rate reflect the errors of exponential fitting of lifetime distribution. The solid line is the least-squares fit to the data with a power law, $R \propto I^N$, where $N = 1.02 \pm 0.08$, indicating a one-electron process. **c**, Current-bias relationship of Na⁺·3D₂O with a Cl⁻ tip at $d = 2$, where the current jumps occur at V_{eff} . **d**, Lateral distance dependence of the positive (red) and negative (black) V_{eff} for Na⁺·3D₂O with a Cl⁻ tip. **e**, Comparison of V_{eff} for Na⁺· n D₂O ($n = 1-5$) at $d = 2, 3$ and 4. Error bars in **d** and **e** reflect the standard deviation (up to 8 different datasets). The tip height in **b-e** is -165 pm referenced to the STM set point on NaCl (100 mV, 10 pA).

is located at the top Cl⁻ site of NaCl, in contrast to the bridge site in the initial state. The conversion barrier between the initial and metastable states is only about 50 meV. The three H₂O molecules of Na⁺·3H₂O cannot simultaneously satisfy an optimal adsorption configuration either at the bridge site or at the top Cl⁻ site owing to the symmetry mismatch with the tetragonal NaCl(001) surface. However, the structure of Na⁺· n H₂O ($n = 1, 2, 4, 5$) matches well with NaCl(001), stabilizing the hydrates at particular sites (bridge or top sites).

Furthermore, we found that the diffusion of Na⁺·3H₂O is coupled with a collective rotation of water in the metastable state (T6 to T11 in Supplementary Fig. 10). Such a rotation requires the water to make only slight adjustments and break minimal bonds with NaCl, leading to a small barrier (about 80 meV). However, removing one H₂O molecule from Na⁺·3H₂O may greatly increase the travelling distance of water during the rotation, while adding one H₂O molecule may block the rotational degrees of freedom owing to the perfect symmetry match between Na⁺·4H₂O and NaCl(001). Therefore, it is the degree of symmetry match between the hydrates and the surface that makes the diffusion barrier sensitive to the number of the water molecules.

These STM/AFM experiments were performed only at low temperature (5 K) and the calculated diffusion barriers correspond to the ones at 0 K. To investigate the surface diffusion at finite temperatures, especially close to room temperature, we carried out classical molecular dynamics simulations (Methods). Figure 4a shows the x - y diffusion trajectories of Na⁺· n H₂O ($n = 1-5$) during a period of 200 ns at 300 K, showing an extraordinarily high mobility of Na⁺·3H₂O (also see Supplementary Videos 1-5). In the zoom-in image (Fig. 4b), two different hopping behaviours were observed. Na⁺·H₂O and Na⁺·2H₂O hop between the bridge sites (Supplementary Videos 6

and 7), while Na⁺·4H₂O and Na⁺·5H₂O hop between the top Cl⁻ sites (Supplementary Videos 9 and 10). Interestingly, Na⁺·3H₂O exhibits a composite behaviour that contains both hopping patterns (Supplementary Video 8). The calculated diffusion mean-square displacements (MSD) at different temperatures are shown in Fig. 4c, revealing that the specific hydration-number effect persists even at room temperature. It is striking that the mobility of Na⁺·3H₂O is more than one order of magnitude larger than that of other clusters at 225 K. We also notice that the diffusion of Na⁺·3H₂O is much faster than that of Na⁺ in bulk solution³⁰.

Supplementary Fig. 11 and Fig. 4d plot the free-energy landscape of different Na⁺ hydrates (Methods). The free-energy minima for Na⁺· n H₂O ($n = 1, 2$) and Na⁺· n H₂O ($n = 4, 5$) are located at the bridge sites and the top Cl⁻ sites, respectively (Supplementary Fig. 11a-d). By contrast, the free-energy surface of Na⁺·3H₂O shows local minima at the Cl⁻ sites in addition to the global minima at the bridge sites (Fig. 4d); these local minima can greatly facilitate the diffusion by truncating the barrier (Supplementary Fig. 11i). From the density distributions of Na⁺ and H₂O of the most stable and metastable Na⁺·3H₂O (Fig. 4e and f), we can identify two characteristic triangular structures (see black dashed triangles and insets, and Supplementary Video 8), closely resembling the initial/final and transition (T3) states obtained by DFT (Fig. 3b). The triangular structures of metastable Na⁺·3H₂O are distributed in four equivalent configurations (see the grey dashed triangles in Fig. 4f), arising from the small rotational energy barrier (about 80 meV) of the three water molecules around the Na⁺, which is much lower than the translational barrier (about 220 meV) (see Supplementary Fig. 9).

More generally, the specific hydration-number effect observed in this work may also exist for other salt ions (Li⁺, K⁺, Cl⁻, and so

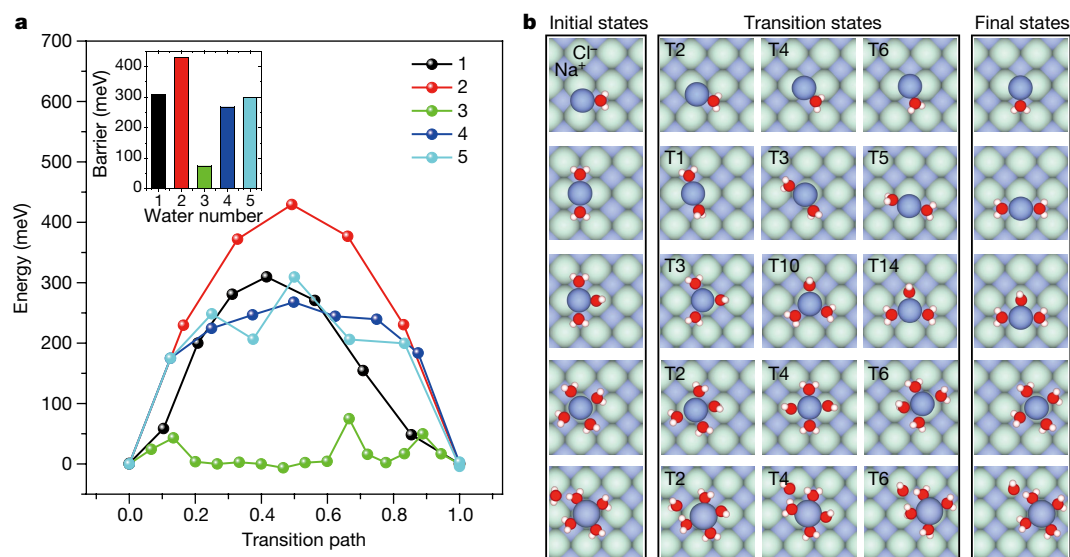


Fig. 3 | Calculated diffusion barrier of Na^+ hydrates by DFT.

a, Energy profiles for the diffusion of $\text{Na}^+ \cdot n\text{H}_2\text{O}$ ($n = 1-5$). The diffusion barriers are compared in the inset. **b**, Snapshots of Na^+ hydrates along the transition path. The first column, the middle three columns and the

last column represent the initial state, transition states and final state of $\text{Na}^+ \cdot n\text{H}_2\text{O}$ ($n = 1-5$), respectively. The number m in the T_m labels at the top left of the images corresponds to the $(m + 1)$ th data point in **a**.

on), but the hydration number can be different depending on the size and the hydration asymmetry of positive and negative ions (Supplementary Fig. 12). Therefore, our results point out a new way to control the ion transport in nanofluidic systems by interfacial

symmetry engineering^{6,9,10}. The techniques developed in this work can easily be extended to different ions and other hydration systems, opening up the possibility of studying various hydration processes down to atomic scale.

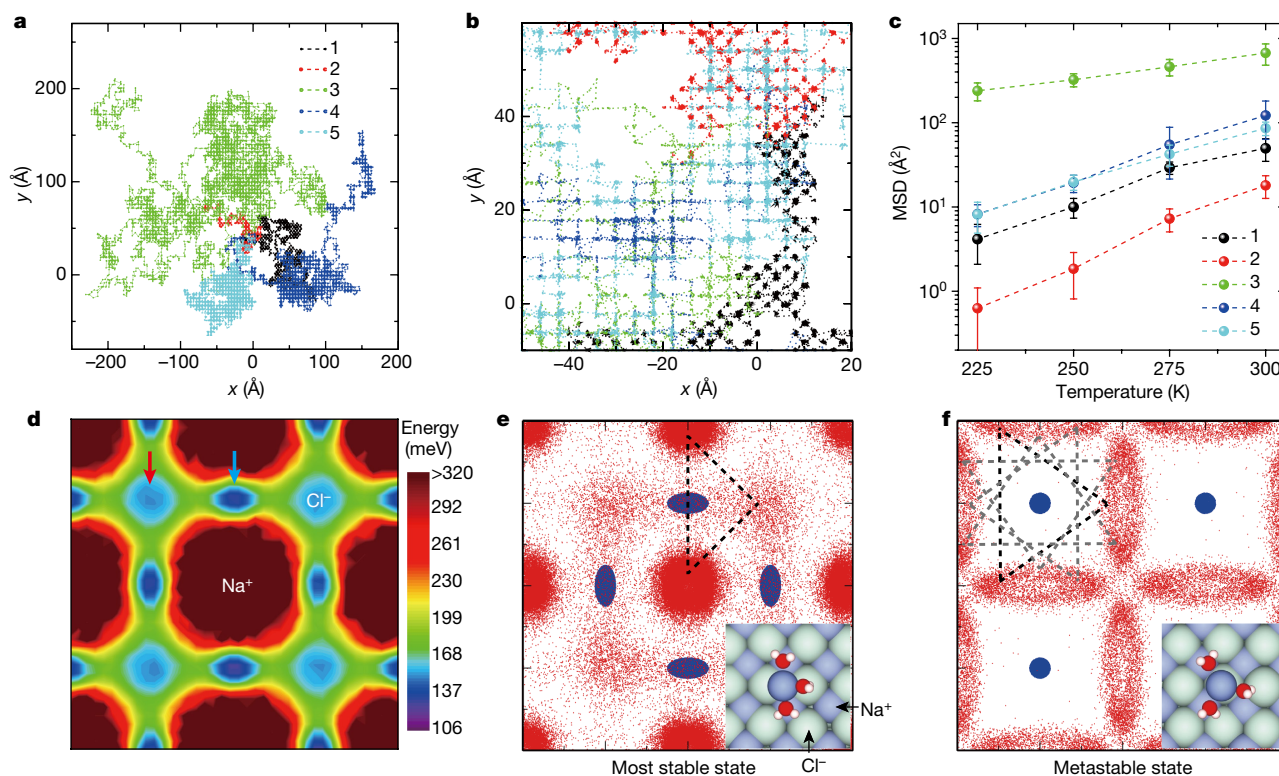


Fig. 4 | Molecular dynamics simulations of the diffusion of Na^+ hydrates at high temperatures. **a**, x - y trajectories of $\text{Na}^+ \cdot n\text{H}_2\text{O}$ ($n = 1-5$) during a period of 200 ns at 300 K. **b**, Zoom-in image of **a** showing different diffusion behaviours. The positions of Na^+ in two consecutive steps are connected by dotted lines. **c**, MSD in 1 ns of $\text{Na}^+ \cdot n\text{H}_2\text{O}$ ($n = 1-5$) between 225 K and 300 K. Error bars reflect the standard deviation from ten different datasets. **d**, The free-energy landscape experienced by $\text{Na}^+ \cdot 3\text{H}_2\text{O}$. It shows global minima at bridge sites (blue arrow) and local minima at top Cl^- sites (red arrow). The positions of Na^+ and Cl^- of the underlying $\text{NaCl}(001)$ surface are labelled. **e**, **f**, Density distributions of

the most stable and metastable $\text{Na}^+ \cdot 3\text{H}_2\text{O}$ on the same area of $\text{NaCl}(001)$ as **d** at 300 K. The blue and red dots represent the positions of Na^+ in the hydrate and of the O atom in H_2O , respectively. The position of Na^+ is constrained within an elliptical area centred at the bridge site (semi-major axis 0.5 Å and semi-minor axis 0.25 Å, **e**) and a circular area centred at the top Cl^- site (radius 0.25 Å, **f**). For clarity, the three water molecules within a representative $\text{Na}^+ \cdot 3\text{H}_2\text{O}$ molecule are connected with black dashed or grey lines. Insets of **e** and **f** are snapshots of $\text{Na}^+ \cdot 3\text{H}_2\text{O}$ at the bridge and top Cl^- sites, respectively. The images in **d-f** are 0.8 nm × 0.8 nm; the insets in **e** and **f** are 1.2 nm × 1.2 nm.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0122-2>.

Received: 18 September 2017; Accepted: 5 March 2018;

Published online: 14 May 2018

- Sipilä, M. et al. Molecular-scale evidence of aerosol particle formation via sequential addition of HIO_3 . *Nature* **537**, 532–534 (2016).
- Klimeš, J., Bowler, D. R. & Michaelides, A. Understanding the role of ions and water molecules in the NaCl dissolution process. *J. Chem. Phys.* **139**, 234702 (2013).
- Cohen-Tanugi, D. & Grossman, J. C. Water desalination across nanoporous graphene. *Nano Lett.* **12**, 3602–3608 (2012).
- Payandeh, J., Scheuer, T., Zheng, N. & Catterall, W. A. The crystal structure of a voltage-gated sodium channel. *Nature* **475**, 353–358 (2011).
- Gouaux, E. & MacKinnon, R. Principles of selective ion transport in channels and pumps. *Science* **310**, 1461–1465 (2005).
- Schoch, R. B., Han, J. & Renaud, P. Transport phenomena in nanofluidics. *Rev. Mod. Phys.* **80**, 839–883 (2008).
- Guo, W., Tian, Y. & Jiang, L. Asymmetric ion transport through ion-channel-mimetic solid-state nanopores. *Acc. Chem. Res.* **46**, 2834–2846 (2013).
- Whitby, M. & Quirke, N. Fluid flow in carbon nanotubes and nanopipes. *Nat. Nanotechnol.* **2**, 87–94 (2007).
- Stein, D., Kruithof, M. & Dekker, C. Surface-charge-governed ion transport in nanofluidic channels. *Phys. Rev. Lett.* **93**, 035901 (2004).
- Duan, C. & Majumdar, A. Anomalous ion transport in 2-nm hydrophilic nanochannels. *Nat. Nanotechnol.* **5**, 848–852 (2010).
- Omta, A. W., Kropman, M. F., Woutersen, S. & Bakker, H. J. Negligible effect of ions on the hydrogen-bond structure in liquid water. *Science* **301**, 347–349 (2003).
- Heisler, I. A. & Meech, S. R. Low-frequency modes of aqueous alkali halide solutions: glimpsing the hydrogen bonding vibration. *Science* **327**, 857–860 (2010).
- Tielrooij, K. J., Garcia-Araez, N., Bonn, M. & Bakker, H. J. Cooperativity in ion hydration. *Science* **328**, 1006–1009 (2010).
- Carrillo-Tripp, M., Saint-Martin, H. & Ortega-Blake, I. A comparative study of the hydration of Na^+ and K^+ with refined polarizable model potentials. *J. Chem. Phys.* **118**, 7062 (2003).
- Jungwirth, P. & Tobias, D. J. Specific ion effects at the air/water interface. *Chem. Rev.* **106**, 1259–1281 (2006).
- Kumagai, T. et al. H-atom relay reactions in real space. *Nat. Mater.* **11**, 167–172 (2012).
- Carrasco, J., Hodgson, A. & Michaelides, A. A molecular perspective of water at metal interfaces. *Nat. Mater.* **11**, 667–674 (2012).
- Guo, J. et al. Real-space imaging of interfacial water with submolecular resolution. *Nat. Mater.* **13**, 184–189 (2014).
- Maier, S. & Salmeron, M. How does water wet a surface? *Acc. Chem. Res.* **48**, 2783–2790 (2015).
- Shiotari, A. & Sugimoto, Y. Ultrahigh-resolution imaging of water networks by atomic force microscopy. *Nat. Commun.* **8**, 14313 (2017).
- Peng, J. et al. Weakly perturbative imaging of interfacial water with submolecular resolution by atomic force microscopy. *Nat. Commun.* **9**, 122 (2018).
- Meng, X. et al. Direct visualization of concerted proton tunnelling in a water nanocluster. *Nat. Phys.* **11**, 235–239 (2015).
- Fukuma, T., Ueda, Y., Yoshioka, S. & Asakawa, H. Atomic-scale distribution of water molecules at the mica–water interface visualized by three-dimensional scanning force microscopy. *Phys. Rev. Lett.* **104**, 016101 (2010).
- Ricci, M., Spijker, P. & Voitchovsky, K. Water-induced correlation between single ions imaged at the solid–liquid interface. *Nat. Commun.* **5**, 4400 (2014).
- Giessibl, F. J. Advances in atomic force microscopy. *Rev. Mod. Phys.* **75**, 949–983 (2003).
- Peng, J. et al. Atomic-scale imaging of the dissolution of NaCl islands by water at low temperature. *J. Phys. Condens. Matter* **29**, 104001 (2017).
- Gross, L. et al. The chemical structure of a molecule resolved by atomic force microscopy. *Science* **325**, 1110–1114 (2009).
- Gawronski, H., Carrasco, J., Michaelides, A. & Morgenstern, K. Manipulation and control of hydrogen bond dynamics in adsorbed ice nanoclusters. *Phys. Rev. Lett.* **101**, 136102 (2008).
- Stipe, B. C., Rezaei, M. A. & Ho, W. Single-molecule vibrational spectroscopy and microscopy. *Science* **280**, 1732–1735 (1998).
- Fuentes-Azcatl, R. & Barbosa, M. C. Sodium chloride, NaCl/e: new force field. *J. Phys. Chem. B* **120**, 2460–2470 (2016).

Acknowledgements This work was supported by the National Key R&D Program under grant numbers 2016YFA0300901, 2017YFA0205003, 2016YFA0300903 and 2015CB856801; the National Natural Science Foundation of China under grant numbers 11634001, 11525520, 21573006 and 11290162/A040106; and the Key Research Program of the Chinese Academy of Sciences under grant numbers XDPB08-1 and XDPB08-4. Y.J. acknowledges support by the National Science Fund for Distinguished Young Scholars (grant number 21725302) and the Cheung Kong Young Scholar Program. P.H. and P.J. acknowledge support from the Czech Academy of Sciences project number MSM100101705 and Premium Academiae and GACR project number 18-09914S. J.G. acknowledges support from the National Postdoctoral Program for Innovative Talents. J.P. acknowledges support from the Weng Hongwu Original Research Foundation under grant number WHW201502. We are grateful for the computational resources provided by the TianHe-1A, TianHe II supercomputer, and the High-performance Computing Platform of Peking University. This work is supported in part by Songshan Lake Laboratory for Material Sciences.

Reviewer information Nature thanks P. Asinari and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Y.J. and E.-G.W. designed and supervised the project. J.P. performed the STM/AFM measurements (with J.G. and R.M.). D.C., J.C., X.-Z.L. and L.-M.X. performed ab initio DFT calculations. Z.H., W.J.X. and Y.Q.G. carried out the classical molecular dynamics simulations. P.H. and P.J. carried out the theoretical simulations of the AFM images (in collaboration with D.C. and B.C.). J.P., D.C., Z.H., J.G., W.J.X., X.-Z.L., Y.Q.G., L.-M.X., E.-G.W. and Y.J. analysed the data. Y.J. wrote the manuscript with input from all other authors. The manuscript reflects the contributions of all authors.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information. is available for this paper at <https://doi.org/10.1038/s41586-018-0122-2>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to L.-M.X. or Y.Q.G. or E.-G.W. or Y.J.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

STM/AFM experiments. All the experiments were performed with a combined noncontact AFM/STM system (Createc, Germany) at 5 K using a qPlus sensor equipped with a tungsten (W) tip (spring constant $k_0 \approx 1,800 \text{ N m}^{-1}$, resonance frequency $f_0 = 23.7 \text{ kHz}$, and quality factor $Q \approx 80,000$). The NaCl(001) bilayer film was obtained by thermally evaporating NaCl crystals onto a clean Au(111) surface at room temperature. To reduce the instability of water molecules induced by the vibrational excitation of inelastic electrons, we used deuterated water (D_2O) instead of H_2O in the experiment. The ultrapure D_2O (Sigma Aldrich, hydrogen-depleted) was used and further purified under vacuum by several freeze-and-pump cycles to remove remaining impurities. The D_2O molecules were dosed in situ onto the sample surface at 5 K through a dosing tube. Bias voltage refers to the sample voltage with respect to the tip. All of the STM topographic images and the AFM frequency shift (Δf) images were obtained with the CO-terminated tips in constant-current and constant-height mode, respectively. The CO tip was obtained by positioning the tip over a CO molecule on the NaCl island at a set point of 100 mV and 20 pA, followed by increasing the bias voltage to 200 mV²¹. The controllable manipulation of water molecules was achieved with the Cl^- -terminated tip at the set point $V = 10 \text{ mV}$, $I = 150 \text{ pA}$. The Cl^- tip was prepared by scanning the NaCl surface in closer proximity (below $V = 5 \text{ mV}$ and $I = 2.5 \text{ nA}$) with a bare metal tip until the Cl atom hopped onto the tip apex²². The construction of the Na^+ hydrates was done with the Cl^- tips (for details see Supplementary Fig. 1).

DFT calculations. DFT calculations were performed using the Vienna ab initio simulation package (VASP)^{31,32}. Projector augmented wave pseudopotentials were used with a cut-off energy of 550 eV for the expansion of the electronic wave functions³³. Van der Waals corrections for dispersion forces were considered by using the optB86b-vdW functional^{34,35}. In our study, the system consisted of a bilayer NaCl(001) on top of Au(111) substrate modelled by a four-layer slab if not specifically mentioned. Similar to ref. 18, a (2×2) NaCl(001) unit cell on a slightly deformed $\begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$ superstructure of the Au(111) substrate was constructed as the surface model. The lattice constant for NaCl(001) surface was set to be 3.9 Å which is the same as in the experiment, and the Au(111) substrate was with a residual strain of about 2%. Supercells of this surface model were used to make the error of water-image interaction negligible with Monkhorst-Pack k -point meshes of spacing denser than $2\pi \times 0.0064 \text{ Å}^{-1}$. The thickness of the vacuum slab was larger than 16 Å and the dipole correction was applied along the surface normal direction^{36,37}. The Au substrate and the bottom layer of the NaCl were fixed in simulations. The Na in the hydrates was positively charged with the charge nearly identical to the Na^+ of the NaCl substrate, based on the Bader charge analysis³⁸. Similar to ref. 22, the Cl^- -terminated tip was modelled using a three-layer Au pyramid of a [111] cleaved face with a Cl atom attached at the end. Energy barriers and paths for the diffusion of the hydration clusters were determined using the cNEB method^{39,40}. The geometry optimizations and the cNEB calculations were performed with a force criterion of 0.01 eV Å^{-1} and 0.02 eV Å^{-1} , respectively. The binding energies (E_{bind}) were calculated by subtracting the total energy of the Na^+ hydrates on the NaCl(001)/Au(111) structure from the sum of the energies of the relaxed bare NaCl(001)/Au(111) substrate, the gas phase of Na and the corresponding isolated water molecules in the gas phase (see Supplementary Fig. 13):

$$E_{\text{bind}} = E[(\text{NaCl}(001)/\text{Au}(111))_{\text{relaxed}}] + n \times E[(\text{H}_2\text{O})_{\text{gas}}] + E[(\text{Na})_{\text{gas}}] - E[(\text{NaCl}(001)/\text{Au}(111) + \text{Na}^+ \cdot n\text{H}_2\text{O})_{\text{relaxed}}]$$

Simulations of AFM images. The Δf images were simulated with a molecular mechanics model including the electrostatic force, based on the methods described in refs 41,42. We used the following parameters of the flexible probe-particle tip model: the effective lateral stiffness $k = 0.75 \text{ N m}^{-1}$ and effective atomic radius $R_c = 1.661 \text{ Å}$. We added a quadrupole-like charge distribution at the tip apex to simulate the CO tip^{21,43} for all the AFM simulations. Comparison between the AFM images and theoretical simulations reveals that the key to the ultrahigh-resolution imaging lies in probing the weak high-order electrostatic force between the quadrupole-like CO-terminated tip and the polar water molecules or ions at large tip-sample distances, in clear contrast to traditional high-resolution AFM imaging at close distances where Pauli repulsion dominates²⁷. This weak interaction allows the imaging and structural determination of the weakly bonded hydrates without inducing any disturbance. The input electrostatic potentials of the Na^+ hydrates on the NaCl(001), employed in AFM simulations, was obtained from DFT calculations. Parameters of Lennard-Jones pairwise potentials for all elements are listed in Supplementary Table 1.

Molecular dynamics simulations. All the molecular dynamics simulation results shown in the paper were obtained by using polarizable force field parameters³⁰ that are developed based on molecular dynamics in electronic continuum theory⁴⁴. The model allowed reproduction of a range of physical and chemical properties of sodium chloride, including the density and the surface tension of pure crystal

system, the viscosity, the dielectric constant, and also the diffusion coefficient in solution³⁰.

The polarizable force field we used is based on a pairwise additive potential that includes a Coulombic treatment of the electrostatic interactions and a Lennard-Jones representation of dispersion-attraction and core repulsion. In this formulation, the potential energy E_{ij} between any pair of non-bonded atoms (i and j) in a system composed of the ions and water molecules is usually expressed as the sum of the van der Waals interaction energy E_{vdW} and the Coulombic interaction energy $E_{\text{Coulombic}}$, namely

$$E_{ij} = 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \lambda_i \lambda_j \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}$$

Here, r_{ij} is the distance between the two atoms; q_i and q_j are the point charges of the atoms; ϵ_0 is the permittivity of vacuum; λ_i and λ_j describe the Coulombic polarizable effect of atoms; σ_{ij} and ϵ_{ij} are the distance at which the interparticle potential is zero and the well depth of the Lennard-Jones potential, respectively. Lorentz-Berteloth combination rules⁴⁵ were used to describe the van der Waals interaction between two different kinds of atoms. Our force field parameters are shown in Supplementary Tables 2 and 3.

To test the effect of force field parameters on the simulated results, we also used a non-polarizable force field parameter, where the SPC/E model was used for water⁴⁶ and the ion parameters were taken from Joung and Cheatham⁴⁷. The results show a consistent conclusion with the polarizable force field.

All classical molecular dynamics simulations were performed using the AMBER14 suite of programs⁴⁸. A four-layer NaCl crystal (atom numbers of $18 \times 18 \times 4$) with a (001) surface was built to support the Na^+ hydrates. Each simulation system was first subjected to 5,000 steps of steepest descent energy minimization, followed by 5,000 steps of conjugate gradient optimization. Then, a 100-ps molecular dynamics simulation was performed to heat the system up to the target temperature, followed by a 10-ns-long normal molecular dynamics simulation to further relax the system. After the initial equilibration, we performed 200-ns calculations for each system with a time step of 2 fs. The temperature was controlled using Langevin dynamics with a collision frequency of 1.0 ps^{-1} . Simulation using Nose-Hoover thermostat yields the same results. The bottom layer of NaCl crystal was restrained by the $2,000 \text{ kcal (mol Å}^2)^{-1}$ force constant, and all classical molecular dynamics simulations were carried out with periodic boundary conditions on the crystal plane. The SHAKE algorithm was used to constrain all bonds involving hydrogen atoms⁴⁹. A cut-off of 1.0 nm was used for van der Waals interactions. A long-range dispersion correction based on an analytical integral assuming an isotropic, uniform bulk particle distribution beyond the cut-off was added to the van der Waals energy and pressure⁴⁸.

Owing to the limitations of our computational ability and the inherently stochastic property of diffusion calculated for small numbers of atoms, it is very difficult to obtain accurate diffusion coefficients (D) for different hydrates. By contrast, bulk calculations sample the trajectories of many more molecules in a more homogeneous environment compared to the current calculations of a cluster on the crystal surface, and are thus faster in yielding converged D values. Instead, we simply used MSD for every nanosecond (up to 20 ns) to compare the mobility of different hydrates. We took the average of MSD from ten different sets of 20-ns data and the error bar reflects the standard deviation.

The equilibrium fractional population distribution of Na^+ hydrates at different sites follows a Boltzmann distribution. The equilibrium ratio of state i is

$$\frac{N_i}{N_{\text{total}}} = \frac{e^{-E_i/RT}}{\sum_{k=1}^{N_{\text{total}}} e^{-E_k/RT}}$$

where e is Euler's constant and E_i is the relative energy of the i th state to the minimum energy state. R is the molar ideal gas constant and T is the temperature. At room temperature, the configurations of hydrates were fully sampled. Thus we used the equilibrium ratio to calculate the free energy landscape of different states, which is

$$\Delta G = -RT \ln \frac{N_i}{N_{\text{total}}}$$

The water orientational time correlation functions $C_2(t)$ were calculated as $C_2(t) = \langle P_2[\mu_{\text{wat}}(0) \cdot \mu_{\text{wat}}(t)] \rangle$, where P_2 is the second-order Legendre polynomial, and $\mu_{\text{wat}}(t)$ is the direction vector of the water dipole at time t .

Data availability. The data that support the findings of this study are available from the corresponding author on reasonable request.

31. Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47**, 558–561 (1993).

32. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
33. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).
34. Klimeš, J., Bowler, D. R. & Michaelides, A. Chemical accuracy for the van der Waals density functional. *J. Phys. Condens. Matter* **22**, 022201 (2010).
35. Klimeš, J., Bowler, D. R. & Michaelides, A. Van der Waals density functionals applied to solids. *Phys. Rev. B* **83**, 195131 (2011).
36. Neugebauer, J. & Scheffler, M. Adsorbate-substrate and adsorbate-adsorbate interactions of Na and K adlayers on Al(111). *Phys. Rev. B* **46**, 16067–16080 (1992).
37. Makov, G. & Payne, M. C. Periodic boundary conditions in ab initio calculations. *Phys. Rev. B* **51**, 4014–4022 (1995).
38. Henkelman, G., Arnaldsson, A. & Jonsson, H. A fast and robust algorithm for Bader decomposition of charge density. *Comput. Mater. Sci.* **36**, 354–360 (2006).
39. Henkelman, G. & Jonsson, H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.* **113**, 9978–9985 (2000).
40. Henkelman, G., Uberuaga, B. P. & Jonsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.* **113**, 9901–9904 (2000).
41. Hapala, P., Temirov, R., Tautz, F. S. & Jelinek, P. Origin of high-resolution IETS-STM images of organic molecules with functionalized tips. *Phys. Rev. Lett.* **113**, 226101 (2014).
42. Hapala, P. et al. Mechanism of high-resolution STM/AFM imaging with functionalized tips. *Phys. Rev. B* **90**, 085421 (2014).
43. Ellner, M. et al. The electric field of CO tips and its relevance for atomic force microscopy. *Nano Lett.* **16**, 1974–1980 (2016).
44. Leontyev, I. V. & Stuchebrukhov, A. A. Polarizable molecular interactions in condensed phase and their equivalent nonpolarizable models. *J. Chem. Phys.* **141**, 014103 (2014).
45. Hansen, J. P. & McDonald, I. R. *Theory of Simple Liquids* 3rd edn (Academic Press, 2006).
46. Berendsen, H. J. C., Grigera, J. R. & Straatsma, T. P. The missing term in effective pair potentials. *J. Phys. Chem.* **91**, 6269–6271 (1987).
47. Joung, I. S. & Cheatham, T. E. Molecular dynamics simulations of the dynamic and energetic properties of alkali and halide ions using water-model-specific ion parameters. *J. Phys. Chem. B* **113**, 13279–13290 (2009).
48. Case, D. A. et al. *AMBER version 14*. <http://ambermd.org> (University of California, San Francisco, 2014).
49. Ryckaert, J. P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).

The origin of squamates revealed by a Middle Triassic lizard from the Italian Alps

Tiago R. Simões^{1*}, Michael W. Caldwell^{1,2}, Mateusz Talanda³, Massimo Bernardi^{4,5}, Alessandro Palci⁶, Oksana Vernygora¹, Federico Bernardini^{7,8}, Lucia Mancini⁹ & Randall L. Nydam¹⁰

Modern squamates (lizards, snakes and amphisbaenians) are the world's most diverse group of tetrapods along with birds¹ and have a long evolutionary history, with the oldest known fossils dating from the Middle Jurassic period—168 million years ago^{2–4}. The evolutionary origin of squamates is contentious because of several issues: (1) a fossil gap of approximately 70 million years exists between the oldest known fossils and their estimated origin^{5–7}; (2) limited sampling of squamates in reptile phylogenies; and (3) conflicts between morphological and molecular hypotheses regarding the origin of crown squamates^{6,8,9}. Here we shed light on these problems by using high-resolution microfocus X-ray computed tomography data from the articulated fossil reptile *Megachirella wachtleri* (Middle Triassic period, Italian Alps¹⁰). We also present a phylogenetic dataset, combining fossils and extant taxa, and morphological and molecular data. We analysed this dataset under different optimality criteria to assess diapsid reptile relationships and the origins of squamates. Our results re-shape the diapsid phylogeny and present evidence that *M. wachtleri* is the oldest known stem squamate. *Megachirella* is 75 million years older than the previously known oldest squamate fossils, partially filling the fossil gap in the origin of lizards, and indicates a more gradual acquisition of squamatan features in diapsid evolution than previously thought. For the first time, to our knowledge, morphological and molecular data are in agreement regarding early squamate evolution, with geckoes—and not iguanians—as the earliest crown clade squamates. Divergence time estimates using relaxed combined morphological and molecular clocks show that lepidosaurs and most other diapsids originated before the Permian/Triassic extinction event, indicating that the Triassic was a period of radiation, not origin, for several diapsid lineages.

Megachirella preserves traits that indicate that it is a lepidosaurian reptile, such as the presence of a well-developed quadrate conch, an ectepicondylar foramen in the humerus and pleurodont dentition. Some of these features led previous authors to recognize the lepidosauromorph affinities of *Megachirella*, which was previously considered as a non-squamate lepidosauromorph, although no definitive conclusions on its phylogenetic placement had ever been reached^{5,10,11}. Yet, the unique condition of *Megachirella* as one of the very few articulated and well-preserved Triassic lepidosauromorphs hints at its potential to help resolve important aspects of lepidosaur evolution. Here we provide substantial new information on *Megachirella*, based on personal observations and high-resolution microfocus X-ray computed tomography (micro-CT) scans, which reveal several previously unnoticed features in *Megachirella* (Fig. 1, Extended Data Figs. 1, 2 and Supplementary Discussion). Results from the micro-CT scans include a combination of features that are found uniquely in squamates: a triradiate squamosal (not tetradiate as in most other diapsids, including rhynchocephalians); the squamosal lacks an anteriorly concave articulatory facet

for the postorbital; a well-developed alar process of the prootic; a well-developed radial condyle on the humerus; an ulnar patella; a secondary curvature of the clavicles; and an expanded epiphysis of the first metacarpal along with the absence of the first distal carpal (suggesting its fusion with the first metacarpal, as observed in modern squamates¹²). Finally, the micro-CT scans indicate that *Megachirella* has features that are absent in all rhynchocephalians (the sister lineage to squamates), including the earliest forms such as *Gephyrosaurus*: the presence of a splenial; the ectopterygoids are directed anteriorly (not laterally as in rhynchocephalians); the presacral pleurocentra lack a notochordal canal; and dorsal (coronoid) expansion of the surangular and dentary bones is absent. The new information presented here, along with our extensive revision of diapsid and early squamate phylogeny, unambiguously resolves the placement of *Megachirella* as the oldest known squamate. As expected for a squamate that is 85 million years (Myr) older than the oldest previously known articulated squamates for which the osteology is well known—*Eichstaettisaurus* and *Ardeosaurus* from the Late Jurassic of Germany^{8,13}—*Megachirella* retains numerous plesiomorphic features. These features are observed in other diapsid reptiles, and some are retained in rhynchocephalians, but they are almost entirely lost in crown squamates. These include amphicoelic vertebrae (although present in geckoes and *Huehuecuetzpalli*), a small quadratojugal, gastralium and an entepicondylar foramen in the humerus.

Assessing the phylogenetic position of *Megachirella* and other lepidosauromorph reptiles is challenging because there has never been a phylogenetic dataset comprising a rich sampling of both non-lepidosaurian diapsid reptiles and squamates. Almost invariably, broad-scale reptile phylogenies have represented the nearly 10,000 extant species and the hundreds of fossil species of squamates as a single operational taxonomic unit^{14–16} (for more examples, see Supplementary Methods). This approach oversimplifies the enormous diversity of phenotypes and genotypes in squamates. Conversely, studies focused on squamate phylogeny never include more than a few taxa outside the Squamata to serve as outgroups^{9,17}. Here we create the first morphological phylogenetic dataset comprising all the main branches of the diapsid tree of life, including extant taxa and fossils from all major lineages of rhynchocephalians (for example, tuataras) and squamates at the species level (Supplementary Data 1–5). We also focused on primary data collection, personally observing numerous specimens covering 100% of the taxa included in this dataset. We performed a meticulous revision of reptile and squamate phylogenetic characters (and created new characters) to avoid issues caused by logical or biological biases in morphological characters¹⁸. Owing to the rich sampling of extant squamate species, we also included molecular data from 16 loci (13 nuclear and 3 mitochondrial). The analyses performed include morphological and combined evidence (morphological and molecular data) analyses of diapsid and lepidosaurian relationships, carried out under multiple phylogenetic inference methods (see Methods).

¹Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada. ²Department of Earth and Atmospheric Sciences, University of Alberta, Edmonton, Alberta, Canada.

³Department of Palaeobiology and Evolution, Faculty of Biology, Biological and Chemical Research Centre, University of Warsaw, Warsaw, Poland. ⁴MUSE, Museo delle Scienze di Trento, Trento, Italy. ⁵School of Earth Sciences, University of Bristol, Bristol, UK. ⁶College of Science and Engineering, Flinders University, Adelaide, South Australia, Australia. ⁷Museo Storico della Fisica e Centro di Studi e Ricerche Enrico Fermi, Roma, Italy. ⁸The Abdus Salam International Centre for Theoretical Physics, Trieste, Italy. ⁹Elettra Sincrotrone Trieste S.C.p.A., Trieste, Italy. ¹⁰Department of Anatomy, Arizona College of Osteopathic Medicine, Midwestern University, Glendale, AZ, USA. *e-mail: tsimoes@ualberta.ca

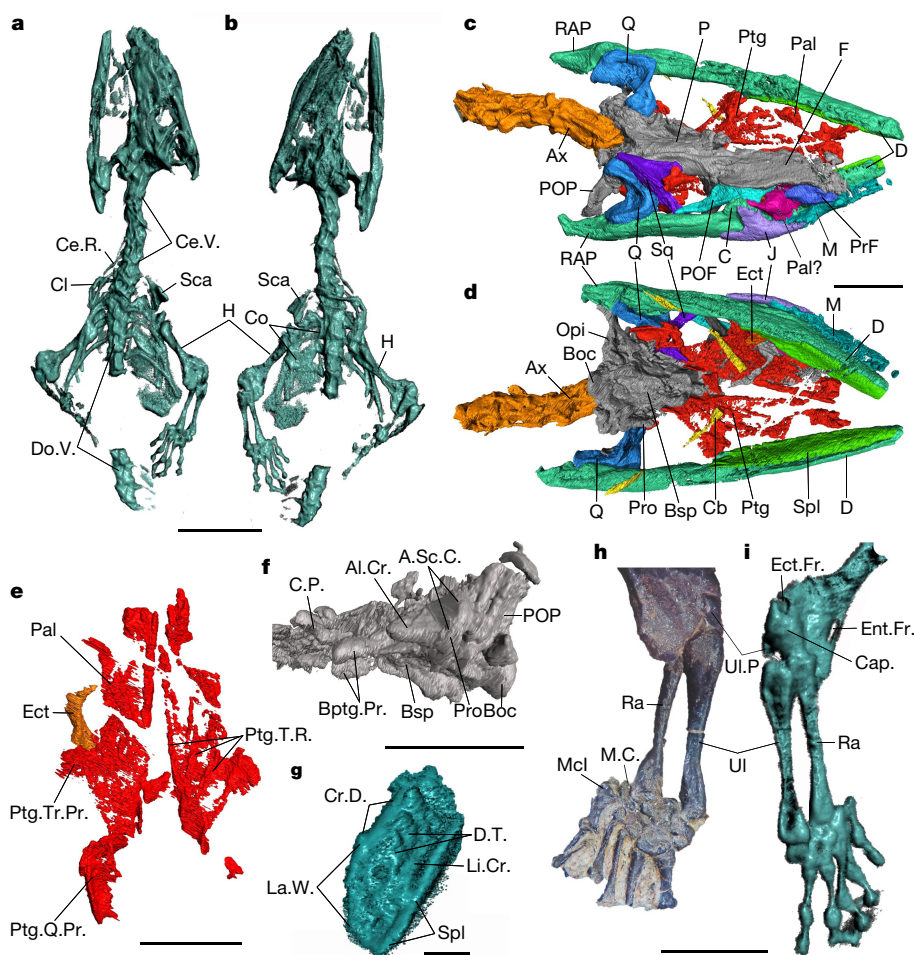


Fig. 1 | Holotype of *M. wachtleri* (PZO 628). **a, b**, Whole skeleton dorsal and ventral views. **c, d**, Skull in dorsal (**c**) and ventral (**d**) views. **e**, Palatal region in ventral view. **f**, Braincase in left lateral view. **g**, Dentary in cross-section. **h, i**, Right forelimb in dorsal (**h**) and ventral (**i**) views. Abbreviations: Al.Cr., prootic alar crest; A.Sc.C., anterior semicircular canal; Ax, axis; Boc, basioccipital; Bptg.Pr., basipterygoid process; Bsp, basisphenoid; C, coronoid; Cap, capitulum; Cb, ceratobranchial; Ce.R., cervical rib; Ce.V., cervical vertebrae; Cl, clavicle; Co, coracoid; C.P., cultriform process; Cr.D., crista dorsalis; D, dentary; Do.V., dorsal vertebrae; D.T., dentary teeth; Ect, ectopterygoid; Ect.Fr., ectepicondylar

foramen; Ent.Fr., entepicondylar foramen; F, frontal; H, humerus; J, jugal; La.W., labial wall; Li.Cr., lingual crest; M, maxilla; M.C., medial centrale; Mcl, metacarpal I; P, parietal; Opi, opisthotics; Pal, palatine; POE, postorbitofrontal; POP, paraoccipital process; PrF, prefrontal; Pro, prootic; Ptg, pterygoid; Ptg.Q.Pr, pterygoid quadrate process; Ptg.T.R., pterygoid tooth rows; Ptg.Tr.Pr., pterygoid transverse process; Q, quadrate; Ra, radius; RAP, retroarticular process; Sca, scapula; Spl, splenial; Sq, squamosal; Ul, ulna; Ul.P., ulnar patella. Scale bars, 10 mm (**a, b**), 5 mm (**c–f, h, i**) and 1 mm (**g**).

Despite the difference in the datasets used (that is, morphology versus combined evidence) and phylogenetic optimality criteria, all results converge on *Megachirella* representing a stem squamate along with *Marmoretta oxoniensis*, from the Middle Jurassic of Britain, and *Huehuecuetzpalli mixtecus*, from the Early Cretaceous period of Mexico (Fig. 2 and Extended Data Figs. 3–8). This resolution is particularly well supported in the combined evidence analysis, in which *Megachirella* has a leaf stability above the overall mean (Extended Data Fig. 9). In analyses with maximum parsimony, *Sophineta cracoviensis* also falls within the Squamata stem, but this is not recovered in the remaining analyses. This indicates that some taxa previously proposed to be early-evolving lepidosauromorphs (for example, *Megachirella* and *Marmoretta*)^{5,10,11} actually represent the oldest known squamates, partially filling the supposed 70-Myr fossil gap in the early history of the clade. Other taxa also considered to be early lepidosauromorphs by previous studies (for example, kuehneosaurids and *Saurosternon*⁵) are consistently found in our results to be nested in other parts of the diapsid tree outside the Lepidosauromorpha. Additionally, all previous morphology- and molecular-based squamate phylogenies available in the literature disagree with each other concerning the earliest-evolving crown group squamates: iguanians for morphology-based analyses^{17,19}, but dibamids and gekkotans for molecular analyses^{7,20,21} (see also Supplementary

Methods). The results of the combined evidence analyses typically match those of the molecular data alone^{6,9}; however, our results show unprecedented agreement between morphological and molecular data, in placing geckoes instead of iguanians among the earliest-evolving squamates. Iguanians are consistently found further crownward in the tree, nested either with anguimorphs and snakes (clade Toxicofera, Extended Data Figs. 3, 5–8), or with teioids (Extended Data Fig. 4). This unprecedented agreement between molecular and morphological data with regards to the early evolution of squamates might be a consequence of our broad sampling of taxa outside squamates (thus affecting character polarity and branch length parameters) and strict criteria for morphological dataset construction.

Megachirella provides unique insights into the early acquisition of squamatan features, as it is the first unequivocal squamate from the Triassic. *Megachirella*, and also *Huehuecuetzpalli*²², show that features that are commonly attributed to squamates characterize crown squamates, but were not yet present in stem squamates. For instance, *Megachirella* and *Huehuecuetzpalli* still retain amphicoelic vertebrae, an entepicondylar foramen, and lack a ball-like distal epiphysis of the ulna. *Megachirella* further indicates that the loss of the quadratojugal and gastralia occurred within squamates, and not at the point of divergence from rhynchocephalians. The same pattern occurs in

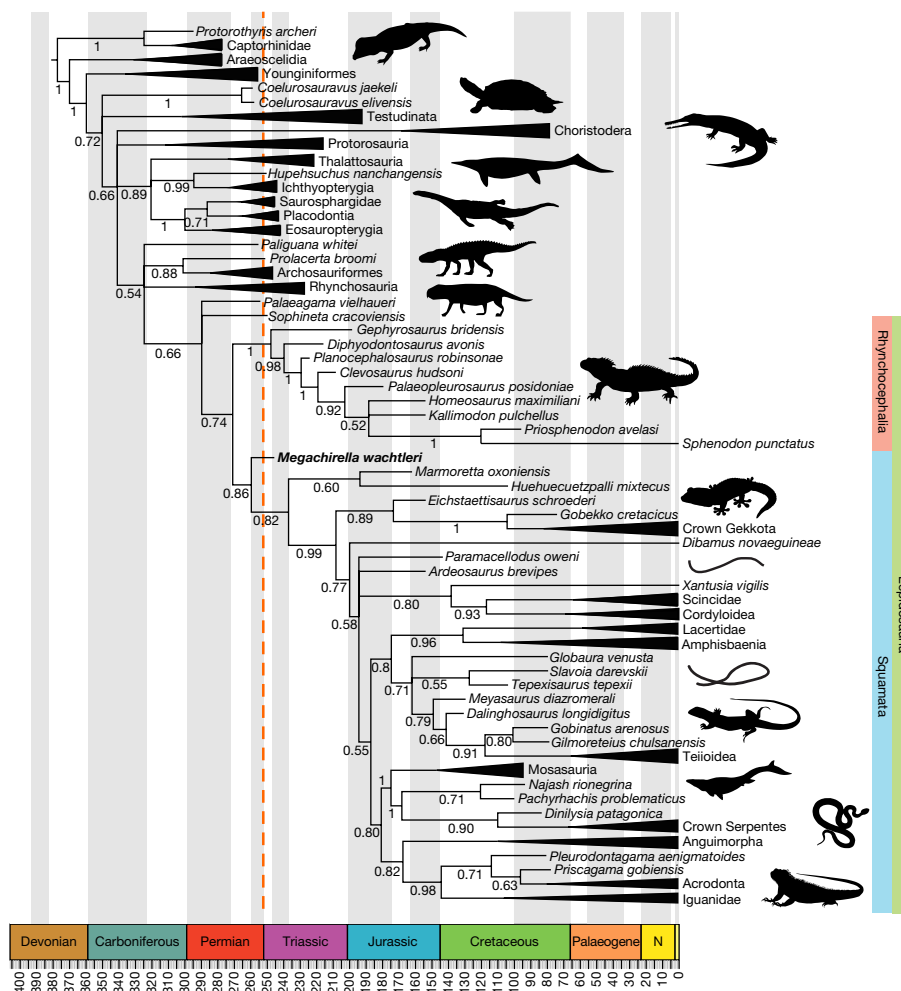


Fig. 2 | Combined evidence relaxed-clock Bayesian inference analysis with total evidence tip and node dating using the fossilized birth–death tree model. Summary of the majority rule consensus tree depicting the median divergence time estimates for the major diapsid and squamate

lineages against a geological time scale. Numbers at nodes indicate posterior probabilities and the orange dashed line represents the Permian/Triassic mass extinction event. For the full tree and 95% highest posterior density on divergence times see Extended Data Fig. 8. N, Neogene period.

rhynchocephalians, for which Triassic and Early Jurassic fossils were previously known²³, and which retain plesiomorphic features (such as the pleurodont dentition) that are absent in most of the later members of that group.

Previous molecular-clock estimates have placed the squamate crown divergence time between the Late Triassic and Early Jurassic^{6,7,24}, and lepidosaurs originating at some point in the Triassic^{5,6} or the Middle Permian period^{7,25}. Our time-calibrated Bayesian inference analyses combine information from both the molecular and morphological relaxed-clocks on lepidosaurs and other diapsid lineages (Fig. 2 and Extended Data Fig. 8), providing a more holistic approach to the divergence time of squamates, lepidosaurs and other diapsids. Our estimates indicate lepidosaurs originated 269 Myr ago (median estimate) in the Middle Permian, and crown squamates 206 Myr ago in the Late Triassic (thus agreeing with recent phylogenomic analyses⁷). Furthermore, our morphological sampling allows a more precise estimate of the origin of the squamate root by the inclusion of fossils now recognized as stem squamates, and thus the age of origin of all squamates can be set at 257 Myr ago, close to the Permian/Triassic mass extinction (PTME).

Some of the oldest known fossils for certain diapsid lineages are known from the earliest Triassic, including ichthyosaurs¹⁶, sauropterygians²⁶ and archosaurs²⁷, with more recent fossil evidence already suggesting the presence of archosauriforms in the Late Permian²⁸, strongly suggesting their divergence preceded the PTME. In accordance, our divergence time estimates for almost all major diapsid lineages (such as lepidosaurs, archosauriforms and marine reptiles) are in the

Permian (Fig. 2 and Extended Data Fig. 8) and not the Triassic (the period from which their oldest known fossils are known). This corresponds to the general expectation that the oldest known fossil of a lineage is likely to be much younger than the actual divergence time for that same lineage²⁹.

The origin of lepidosaurs and other major diapsid lineages before the PTME contradicts previous ideas suggesting that those groups originated in the aftermath of the greatest mass extinction in Earth's history³⁰. Instead, our results indicate those lineages already existed, but radiated in the Triassic. It is likely that the PTME opened new niches and opportunities to lineages previously restricted in diversity, thus enabling their radiation in the Triassic into numerous forms and sizes, occupying all major biomes on the planet.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0093-3>

Received: 5 December 2017; Accepted: 28 February 2018;
Published online 30 May 2018.

1. Uetz, P. & Hošek, J. *The Reptile Database* <http://www.reptile-database.org> (2017).
2. Nessov, L. Late Mesozoic amphibians and lizards of Soviet Middle Asia. *Acta Zool. Cracov.* **31**, 475–486 (1988).

3. Fedorov, P. & Nesson, L. A lizard from the boundary of the Middle and Late Jurassic of north-east Fergana. *Bull. St Petersburg Univ. Geol. Geogr.* **3**, 9–14 (1992).
4. Evans, S. Crown group lizards (Reptilia, Squamata) from the Middle Jurassic of Britain. *Palaeontographica. A* **250**, 123–154 (1998).
5. Jones, M. E. H. et al. Integration of molecules and new fossils supports a Triassic origin for Lepidosauria (lizards, snakes, and tuatara). *BMC Evol. Biol.* **13**, 208 (2013).
6. Pyron, R. A. Novel approaches for phylogenetic inference from morphological data and total-evidence dating in squamate reptiles (lizards, snakes, and amphisbaenians). *Syst. Biol.* **66**, 38–56 (2017).
7. Irisarri, I. et al. Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nat. Ecol. Evol.* **1**, 1370–1378 (2017).
8. Simões, T. R., Caldwell, M. W., Nydam, R. L. & Jiménez-Huidobro, P. Osteology, phylogeny, and functional morphology of two Jurassic lizard species and the early evolution of scansoriality in geckoes. *Zool. J. Linn. Soc.* **180**, 216–241 (2017).
9. Reeder, T. W. et al. Integrated analyses resolve conflicts over squamate reptile phylogeny and reveal unexpected placements for fossil taxa. *PLoS ONE* **10**, e0118199 (2015).
10. Renesto, S. & Posenato, R. A new lepidosauromorph reptile from the Middle Triassic of the Dolomites (Northern Italy). *Riv. Ital. Paleontol. Stratigr.* **109**, 463–474 (2003).
11. Renesto, S. & Bernardi, M. Redescription and phylogenetic relationships of *Megachirella wachtleri* Renesto et Posenato, 2003 (Reptilia, Diapsida). *Palaontol. Z.* **88**, 197–210 (2014).
12. Carroll, R. L. in *Problems in Vertebrate Evolution* Vol. 4 (eds Andrews, S. M. et al.) 1–28 (Academic Press, London, 1977).
13. Mateer, N. Osteology of the Jurassic lizard *Ardeosaurus brevipes* (Meyer). *Palaeontology* **25**, 461–469 (1982).
14. Chen, X. H., Motani, R., Cheng, L., Jiang, D.-Y. & Rieppel, O. The enigmatic marine reptile *Nanchangosaurus* from the lower triassic of Hubei, China and the phylogenetic affinities of Hupehsuchia. *PLoS ONE* **9**, e102361 (2014).
15. Müller, J. in *Recent Advances in the Origin and Early Radiation of Vertebrates* (eds Arratia, G. et al.) 379–408 (F. Pfeil, München, 2004).
16. Motani, R., Minoura, N. & Ando, T. Ichthyosaur relationships illuminated by new primitive skeletons from Japan. *Nature* **393**, 255–257 (1998).
17. Conrad, J. L. Phylogeny and systematics of Squamata (Reptilia) based on morphology. *Bull. Am. Mus. Nat. Hist.* **310**, 1–182 (2008).
18. Simões, T. R., Caldwell, M. W., Palci, A. & Nydam, R. L. Giant taxon-character matrices: quality of character constructions remains critical regardless of size. *Cladistics* **33**, 198–219 (2017).
19. Gauthier, J. A., Kearney, M., Maisano, J. A., Rieppel, O. & Behlke, A. D. B. Assembling the squamate tree of life: perspectives from the phenotype and the fossil record. *Bull. Peabody Mus. Nat. Hist.* **53**, 3–308 (2012).
20. Pyron, R. A., Burbrink, F. T. & Wiens, J. J. A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. *BMC Evol. Biol.* **13**, 93 (2013).
21. Vidal, N. & Hedges, S. B. The phylogeny of squamate reptiles (lizards, snakes, and amphisbaenians) inferred from nine nuclear protein-coding genes. *C. R. Biol.* **328**, 1000–1008 (2005).
22. Reynoso, V.-H. *Huehucuetzpalli mixtecus* gen. et sp. nov.: a basal squamate (Reptilia) from the Early Cretaceous of Tepexi de Rodríguez, Central México. *Phil. Trans. R. Soc. Lond. B* **353**, 477–500 (1998).
23. Evans, S. E. The skull of a new eosuchian reptile from the Lower Jurassic of South Wales. *Zool. J. Linn. Soc.* **70**, 203–264 (1980).
24. Zheng, Y. & Wiens, J. J. Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species. *Mol. Phylogenet. Evol.* **94**, 537–547 (2016).
25. Huggall, A. F., Foster, R. & Lee, M. S. Y. Calibration choice, rate smoothing, and the pattern of tetrapod diversification according to the long nuclear gene RAG-1. *Syst. Biol.* **56**, 543–563 (2007).
26. Jiang, D.-Y. et al. The Early Triassic Eosauroptrygian *Majiashanosaurus discocoracoidis*, gen. et sp. nov. (Reptilia, Saurpterygia), from Chaohu, Anhui Province, People's Republic of China. *J. Vertebr. Paleontol.* **34**, 1044–1052 (2014).
27. Butler, R. J. et al. The sail-backed reptile *Ctenosauriscus* from the latest Early Triassic of Germany and the timing and biogeography of the early archosaur radiation. *PLoS ONE* **6**, e25693 (2011).
28. Bernardi, M., Klein, H., Petti, F. M. & Ezcurra, M. D. The origin and early radiation of archosauriforms: integrating the skeletal and footprint record. *PLoS ONE* **10**, e0128449 (2015).
29. Ho, S. Y. & Phillips, M. J. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. *Syst. Biol.* **58**, 367–380 (2009).
30. Chen, Z.-Q. & Benton, M. J. The timing and pattern of biotic recovery following the end-Permian mass extinction. *Nat. Geosci.* **5**, 375–383 (2012).

Acknowledgements We are grateful for funding from the Vanier Canada and the Izaak Walton Killam Memorial PhD scholarships to T.R.S.; Euregio Science Fund (call 2014, IPN16) to M.B.; Midwestern University Intramural Funds to R.L.N.; Natural Science and Engineering Research Council of Canada Discovery Grant to M.W.C. (23458); Alberta Ukrainian Centennial Scholarship to O.V.; and National Science Centre grant 2014/13/N/NZ8/02467 to M.T.; and E. Kustatscher for access to the holotype of *M. wachtleri*.

Reviewer information Nature thanks M. Baron, J.-C. Rage and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions T.R.S. conducted phylogenetic data collection and analyses; M.W.C. conceived the project; F.B., L.M., M.B., T.R.S. and A.P. conducted micro-CT scans and computed tomography segmentations; T.R.S. and O.V. performed molecular sequence alignment; T.R.S., M.B., M.T., A.P. and R.L.N. performed morphological description; all authors contributed to writing and discussions.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0093-3>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0093-3>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to T.R.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Micro-CT. The holotype of *Megachirella wachleri* was analysed by micro-CT at the Multidisciplinary Laboratory of the Abdus Salam International Centre of Theoretical Physics (Trieste, Italy), using a system specifically designed in collaboration with Elettra-Sincrotrone Trieste (Basovizza, Italy) for the study of palaeontological and archaeological materials³¹. The micro-CT acquisition of the complete specimen was carried out by using a sealed X-ray source (Hamamatsu L8121-03) at a voltage of 150 kV, a current of 100 μ A and with a focal spot size of 20 μ m. The X-ray beam was filtered by a 1.5-mm-thick aluminium absorber. A set of 2,400 projections of the sample were recorded over a total scan angle of 360° by a flat panel detector (Hamamatsu C7942SK-25) with an exposure time of 2.0 s. The resulting micro-CT slices were reconstructed in 16-bit format using the commercial software DigiXCT (DIGISENS) and an isotropic voxel size of 42.51 μ m. Additionally, the proximal part of the sample was re-analysed (voltage 150 kV, current 100 μ A, 1-mm copper filter, exposure time/projection 3.0 s and 1,800 projections over 360°) setting an effective pixel size of 18 μ m and reconstructed using the same software to achieve a higher spatial resolution.

Morphological dataset construction. All taxa used in this study were personally observed by at least one of us, and more than half by two or more of the co-authors. The new dataset presented herein includes a large sample of species of squamates, as well as a broad variety of non-squamatan lepidosaurs and non-lepidosaurian diapsid species, representing all of the major clades of diapsid reptiles. Characters were assessed based on primary homology assessment and according to strict criteria for character construction, to avoid biases owing to logical or biological dependencies across characters, overweighing of any anatomical attributes and many other issues that may affect the morphological component of phylogenetic datasets¹⁸. We selected *Protorothyris archeri* as the outgroup to our analyses and all morphological characters were treated as unordered (see Supplementary Methods for additional details).

Molecular dataset alignment, model selection and partitions. The molecular dataset consists of 16 genetic markers (13 nuclear and 3 mitochondrial loci) for 38 extant taxa. A complete list of sampled loci and sequence lengths is provided in Supplementary Table 1. Sequence data for the selected coding regions were obtained from GenBank (Supplementary Data 2). For three ingroup taxa, *Liolaemus signifer*, *Pristidactylus scapulatus* and *Stenocercus scapularis*, for which molecular data were not available, we used sequences of the congeneric species, *L. ornatus*, *P. torquatus* and *S. guentheri*, respectively. Sequences were aligned in the MAFFT 7.245³² online server using the global alignment strategy with iterative refinement and consistency scores. For the protein-coding genes, alignments were verified by translating nucleotide sequences to amino acids. The final multiple sequence alignment was concatenated and visually examined in Mesquite 3.04³³. Molecular sequences from all extant taxa were analysed for the best partitioning scheme and model of evolution using PartitionFinder2³⁴ under Akaike information criterion.

Equal weights maximum parsimony analysis. Analyses were conducted in TNT v.1.1³⁵ using the new technology search algorithms. This strategy enables the sampling of trees from a broader spectrum of local optima than is allowed by the heuristic search with ratchet runs in PAUP* v.4.0 beta 10, especially for large datasets^{35,36}. Tree searches were conducted using 1,000 initial trees by random addition sequences with 100 iterations or rounds for each of the four NTS algorithms: sectorial search, ratchet, drift and tree fusing. The output trees were used as the starting trees for subsequent runs, using 1,000 iterations/rounds of each of the new technology search algorithms. The latter step was repeated once, and the final output trees were filtered for all the most parsimonious trees (MPTs). A total of 621 MPTs were obtained with 2,268 steps each.

Implied weights maximum parsimony analysis. Analyses were also conducted in TNT, using the implied weighting algorithm³⁷, with a $K = 12$ and collapsing all branches with support = 0. Tree searches were conducted as performed for the equal weights parsimony analysis. Larger K values than the default (3.0) are indicated to perform better for large datasets³⁸. A total of five best fit trees were obtained (fit = 91.768892) and used to calculate the strict consensus tree.

Bayesian inference analyses. Analyses were conducted using Mr. Bayes v.3.2.6³⁹ using the Cedar computer cluster made available through Compute Canada and the CIPRES Science Gateway v.3.3⁴⁰. Molecular partitions were analysed using the models of evolution obtained from PartitionFinder2 (see dataset), and the morphological partition was analysed with the MkV model⁴¹.

The distribution for rate heterogeneity was tested for best fit to the data under both γ and log-normal distributions, as it was recently demonstrated that a log-normal distribution may better fit morphological data for a large variety of datasets^{42,43}. Fit to the data was assessed using Bayes factors $[B_{10}]^{44,45}$ calculated with the marginal model likelihoods obtained from the stepping-stone sampling method⁴⁶. The interpretation of the results of the model fit to the data was performed as previously described⁴⁵: when $2\log_e(B) > 2$ (positive evidence against model M_0);

when $2\log_e(B) > 6$ (strong evidence against model M_0); when $2\log_e(B) > 10$ (very strong evidence against model M_0). However, $2\log_e(B)$ was less than one between the γ and log-normal runs, indicating that there was no significant difference in fit to the morphological data between both distributions. The morphological partition was thus analysed under the γ model for all subsequent analyses.

Time-calibrated relaxed-clock Bayesian inference analyses. We implemented 'total-evidence-dating' using the fossilized birth-death tree model with sampled ancestors, under a relaxed-clock model in Mr. Bayes v.3.2.6^{47–49}. The chosen relaxed-clock model is the independent γ rate relaxed-clock model⁵⁰. This is a continuous uncorrelated relaxed-clock model using a gamma distribution to assess clock rate variation across lineages. The latter is compatible with the fossilized birth-death tree model, unlike the compound Poisson process relaxed-clock model⁴⁸. The base clock rate was given an informative prior, which was derived from the non-clock Bayesian inference analysis: the median value for tree height in substitutions from the entire posterior trees sample divided by the age of the tree, which is based on the median of the distribution for the root prior: $25.1658/325.45 = 0.0773$, in natural log scale = -2.560061 . We chose to use the exponent of the mean to provide a broad standard deviation ($e^{0.0773} = 1.080366$) as previously recommended⁶. The sampling strategy was set to diversity, which is more appropriate when extant taxa are sampled in a manner that maximizes diversity (as performed herein) and fossils are sampled randomly^{47,48}. Diversity sampling is very common in higher-level phylogenies, and not accounting for it has a deep effect on tree inference, pushing divergence times further back and creating unreasonably older and more variable divergence times^{48,51}. This is a considerable advantage of using Mr. Bayes for divergence time estimates over current implementations available in the software package BEAST⁵².

The wealth of fossil taxa in our dataset, including some of the oldest known taxa for many clades, provided numerous calibration points. Therefore, the vast majority of our calibrations were based on tip dating, which accounts for the uncertainty in the placement of fossil taxa and avoids the issue of bound estimates for node-based age calibrations⁴⁷. The fossil ages used for tip dating correspond to the uniform prior distributions on the age range of the stratigraphic occurrence of the fossils (available in Supplementary Table 2). However, it has recently been demonstrated that using tip dates only can contribute to unrealistically older divergence time estimates for some clades^{53,54}. Therefore, when we lacked the oldest known fossils for any of the clades in our analysis (namely, captorhinids, choristoderes, snakes and rhynchocephalians), we used node-age calibrations with a soft lower bound as long as the age of the oldest known fossil was well-established and there was overwhelming support in the literature (and in all our other analyses) for their monophyly. Combined with the diversity sampling strategy, the latter dating protocol can ensure reliable divergence time estimates.

The age of the root was set with a soft lower bound, which gives a low (but non-zero) likelihood of the age being older than the lower bound value. Minimum and maximum root bounds were placed as follows. The minimum age was set at the oldest possible age for the oldest known reptile, *Hylonomus* (from the Joggins Formation in Nova Scotia, Canada), which comes from the late Bashkirian Stage (early Pennsylvanian, Late Carboniferous) and is between 318 and 315 Myr old⁵⁵. Considering *Petrolacosaurus* may be as much as 307 Myr old, placing the minimum age at 318 Myr seems consistent, as the most recent common ancestor of diapsids and captorhinids must have been at least a few million years older than *Petrolacosaurus*. The maximum age was based on the maximum soft age for the reptile-synapsid split⁵⁶, 332.9 Ma.

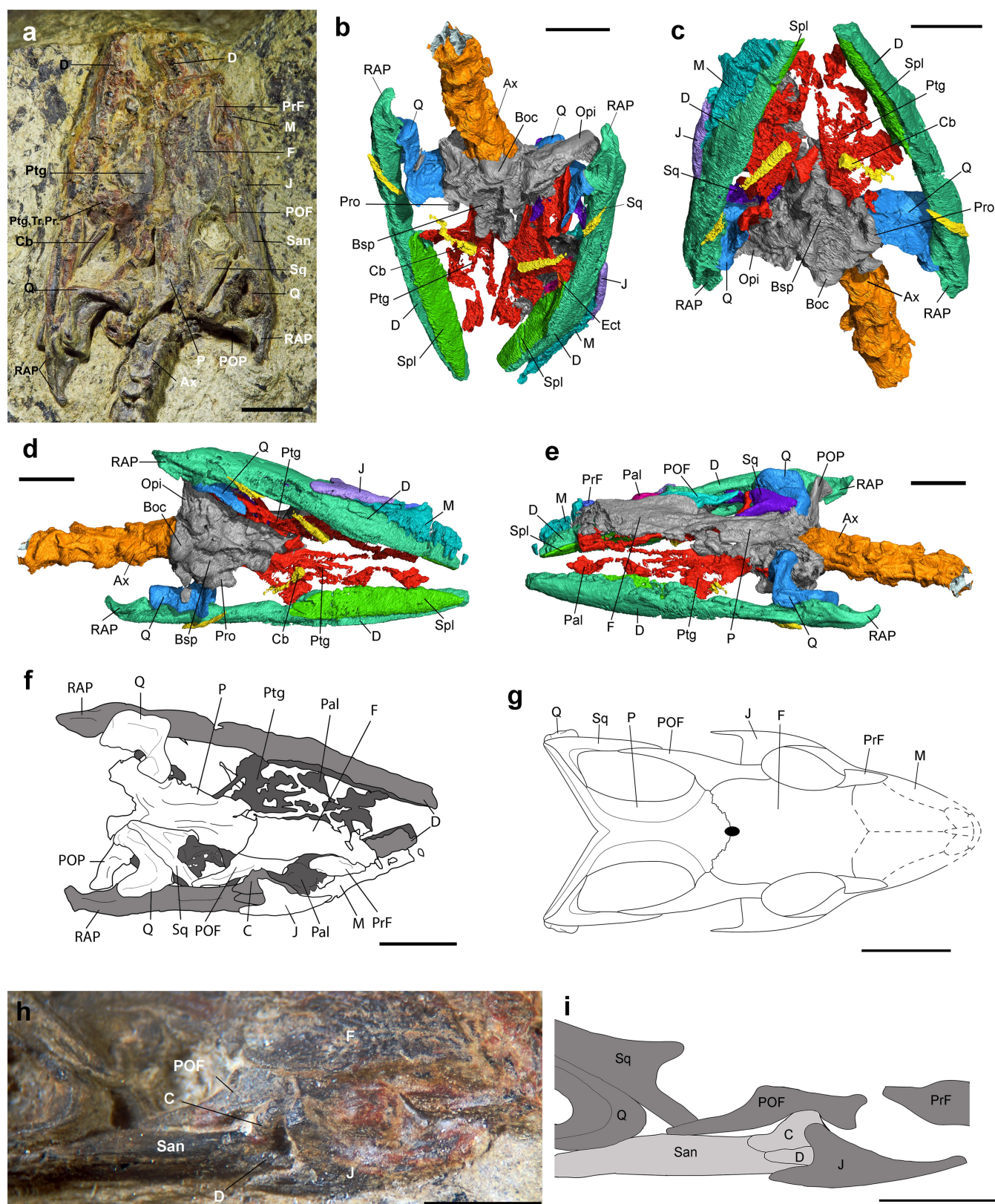
Convergence of independent runs was assessed using an average standard deviation of split frequencies of approximately 0.01, potential scale reduction factors of approximately 1 for all parameters⁵⁷ and an effective sample size greater than 200 for each parameter.

Leaf stability. Leaf stability was assessed using RogueNaRok⁵⁸, which allows assessing the difference between the highest and the second highest support values for alternative resolutions of each taxon quartet or triplet in the dataset (LSdif)⁵⁹. We applied this method to the posterior trees from the Bayesian inference analysis including both the morphological and molecular data. Because of the large number of taxa and large number of trees, it was necessary to downsample the total number of posterior trees from each analysis (100,000 trees after discarding burn-in). The final sample consisted of 10,000 trees (selecting one at every 10 trees) using the Burntrees script for Perl (<https://github.com/nylander/Burntrees>). Taxon names and raw data relating to each number depicted in Extended Data Fig. 9 can be found in Supplementary Table 3.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

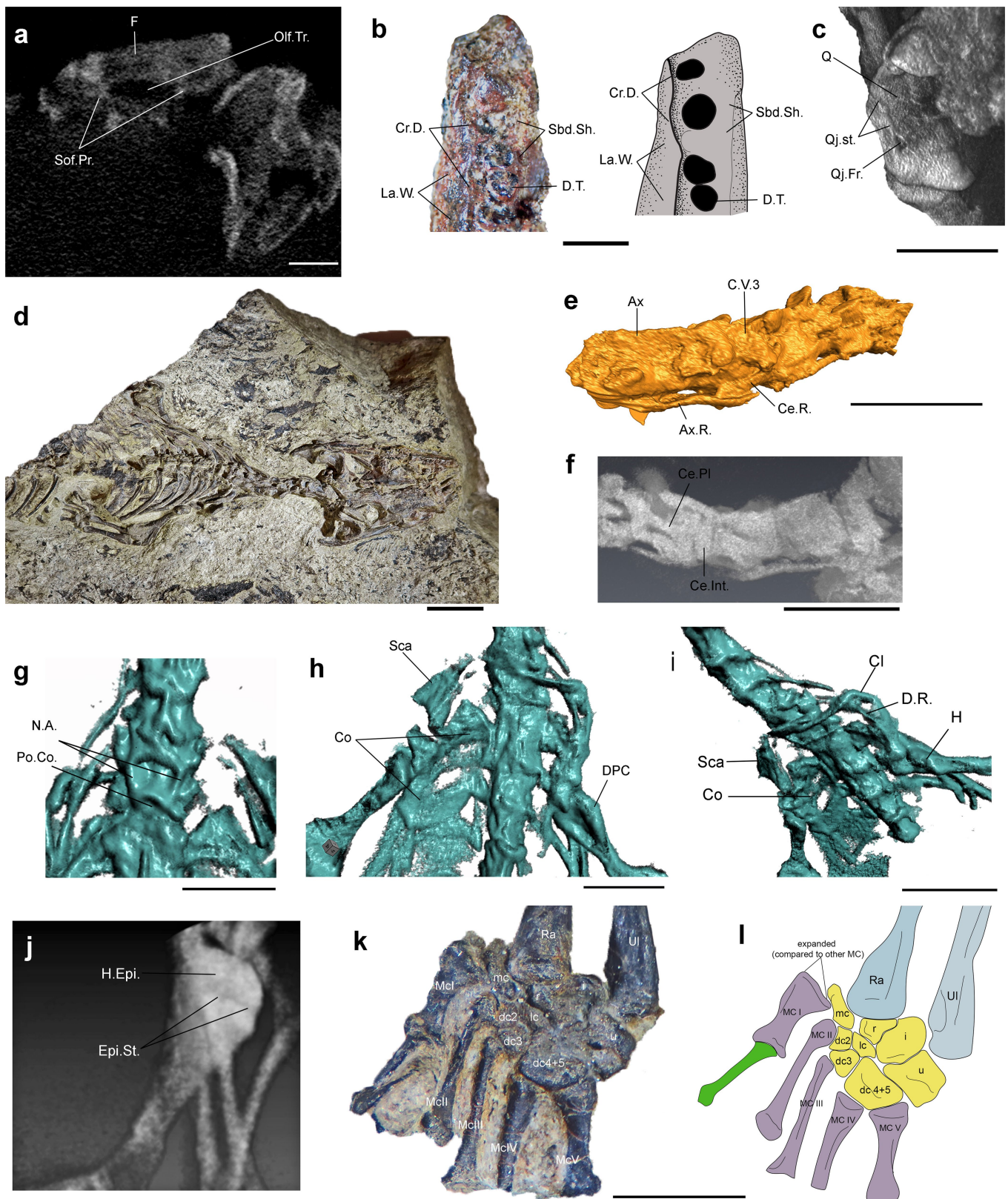
Data availability. The micro-CT scan data are available from the authors upon reasonable request. The morphological and molecular datasets for the phylogenetic analyses, including the Mr. Bayes parameters block, are available as Supplementary Information.

31. Tuniz, C. et al. The ICTP-Elettra X-ray laboratory for cultural heritage and archaeology. *Nucl. Instrum. Methods Phys. Res. A* **711**, 106–110 (2013).
32. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
33. Maddison, W. P. & Maddison, D. R. *Mesquite: a Modular System for Evolutionary Analysis* v.3.04 <http://mesquiteproject.org> (2015).
34. Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T. & Calcott, B. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol. Biol. Evol.* **34**, 772–773 (2017).
35. Goloboff, P. A., Farris, J. S. & Nixon, K. C. TNT, a free program for phylogenetic analysis. *Cladistics* **24**, 774–786 (2008).
36. Goloboff, P. A. Analyzing large data sets in reasonable times: solutions for composite optima. *Cladistics* **15**, 415–428 (1999).
37. Goloboff, P. A., Carpenter, J. M., Arias, J. S. & Esquivel, D. R. M. Weighting against homoplasy improves phylogenetic analysis of morphological data sets. *Cladistics* **24**, 758–773 (2008).
38. Goloboff, P. A., Torres, A. & Arias, J. S. Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology. *Cladistics* <https://doi.org/10.1111/cla.12205> (2017).
39. Ronquist, F. et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012).
40. Miller, M. A., Pfeiffer, W. & Schwartz, T. in *Proceedings of the 2010 Gateway Computing Environments Workshop (GCE)* (IEEE, 2010).
41. Lewis, P. O. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* **50**, 913–925 (2001).
42. Harrison, L. B. & Larsson, H. C. Among-character rate variation distributions in phylogenetic analysis of discrete morphological characters. *Syst. Biol.* **64**, 307–324 (2015).
43. Wagner, P. J. Modelling rate distributions using character compatibility: implications for morphological evolution among fossil invertebrates. *Biol. Lett.* **8**, 143–146 (2012).
44. Nylander, J. A. A., Ronquist, F., Huelsenbeck, J. P. & Nieves-Aldrey, J. L. Bayesian phylogenetic analysis of combined data. *Syst. Biol.* **53**, 47–67 (2004).
45. Kass, R. E. & Raftery, A. E. Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
46. Xie, W., Lewis, P. O., Fan, Y., Kuo, L. & Chen, M.-H. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.* **60**, 150–160 (2011).
47. Ronquist, F. et al. A total-evidence approach to dating with fossils, applied to the early radiation of the hymenoptera. *Syst. Biol.* **61**, 973–999 (2012).
48. Zhang, C., Stadler, T., Klopstein, S., Heath, T. A. & Ronquist, F. Total-evidence dating under the fossilized birth–death process. *Syst. Biol.* **65**, 228–249 (2016).
49. Stadler, T. Sampling-through-time in birth–death trees. *J. Theor. Biol.* **267**, 396–404 (2010).
50. Lepage, T., Bryant, D., Philippe, H. & Lartillot, N. A general comparison of relaxed molecular clock models. *Mol. Biol. Evol.* **24**, 2669–2680 (2007).
51. Höhna, S., Stadler, T., Ronquist, F. & Britton, T. Inferring speciation and extinction rates under different sampling schemes. *Mol. Biol. Evol.* **28**, 2577–2589 (2011).
52. Bouckaert, R. et al. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **10**, e1003537 (2014).
53. O'Reilly, J. E., Dos Reis, M. & Donoghue, P. C. J. Dating tips for divergence-time estimation. *Trends Genet.* **31**, 637–650 (2015).
54. O'Reilly, J. E. & Donoghue, P. C. Tips and nodes are complementary not competing approaches to the calibration of molecular clocks. *Biol. Lett.* **12**, 20150975 (2016).
55. Ogg, J. G., Ogg, G. & Gradstein, F. M. *A Concise Geologic Time Scale* (Elsevier, Amsterdam, 2016).
56. Benton, M. J. et al. Constraints on the timescale of animal evolutionary history. *Palaeontol. Electronica* **18**, 1–106 (2015).
57. Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–472 (1992).
58. Aberer, A. J., Krompass, D. & Stamatakis, A. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Syst. Biol.* **62**, 162–166 (2013).
59. Wilkinson, M. Identifying stable reference taxa for phylogenetic nomenclature. *Zool. Scr.* **35**, 109–112 (2006).



Extended Data Fig. 1 | Cranial anatomy of *M. wachtleri* (PZO 628) based on personal examination and micro-CT scan data. a, Skull in dorsal view. b, Skull in posteroventral view. c, Skull in anteroventral view. d, Skull in right ventrolateral view. e, Skull in left dorsal lateral view. f, Line

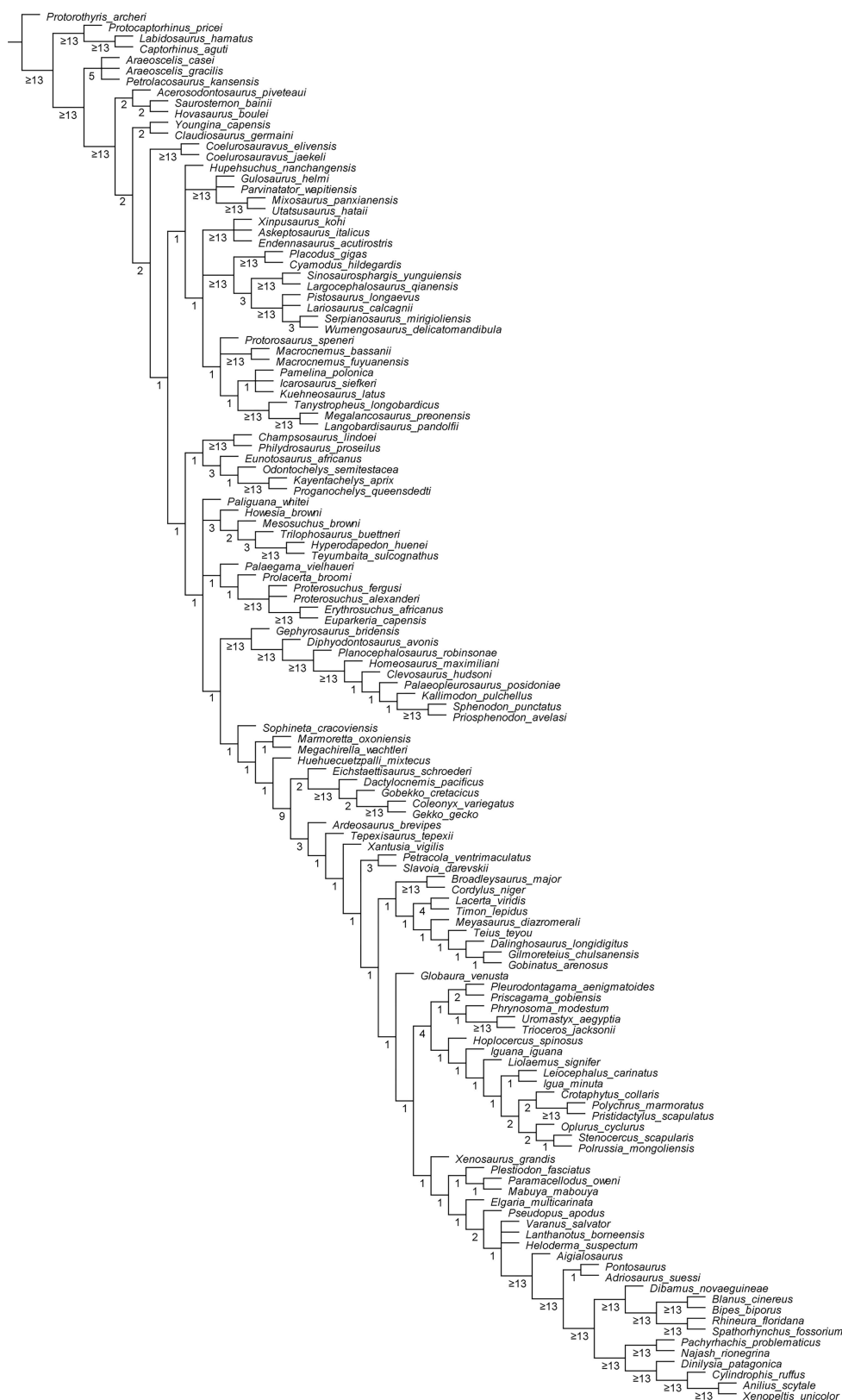
drawing of the skull in dorsal view. g, Reconstruction of the skull in dorsal view. h, Detailed view of right lateral side of the skull. i, Drawing of the view in h. San, surangular. Scale bars, 5 mm (a–g).



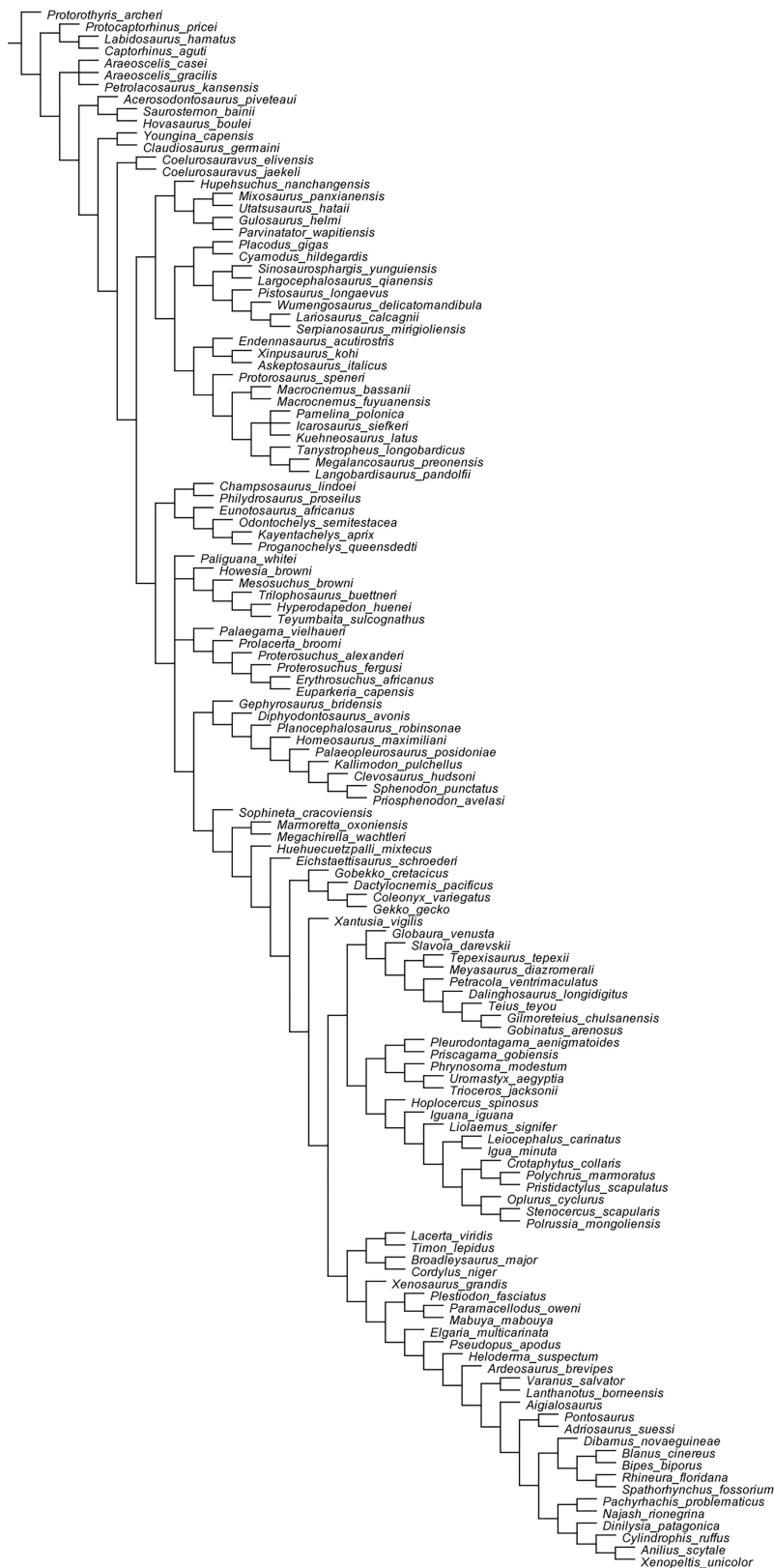
Extended Data Fig. 2 | Cranial and postcranial anatomy of *M. wachtleri* (PZO 628) based on personal examination and micro-CT scan data.

a, Cross-section of the skull at the level of the frontals in anterior view. **b**, Details of the anterior end of the left dentary in occlusal view. **c**, Left quadrate. **d**, Whole body of the holotype as preserved in the slab (dorsal view). **e**, Anterior cervical vertebrae in left lateral view. **f**, Longitudinal section of the anterior cervicals in ventral view. **g**, Last cervicals and anterior dorsals in dorsal view. **h**, Pectoral girdle in ventral view. **i**, Pectoral girdle in left ventrolateral view. **j**, Right humerus in ventral view. **k**, Right

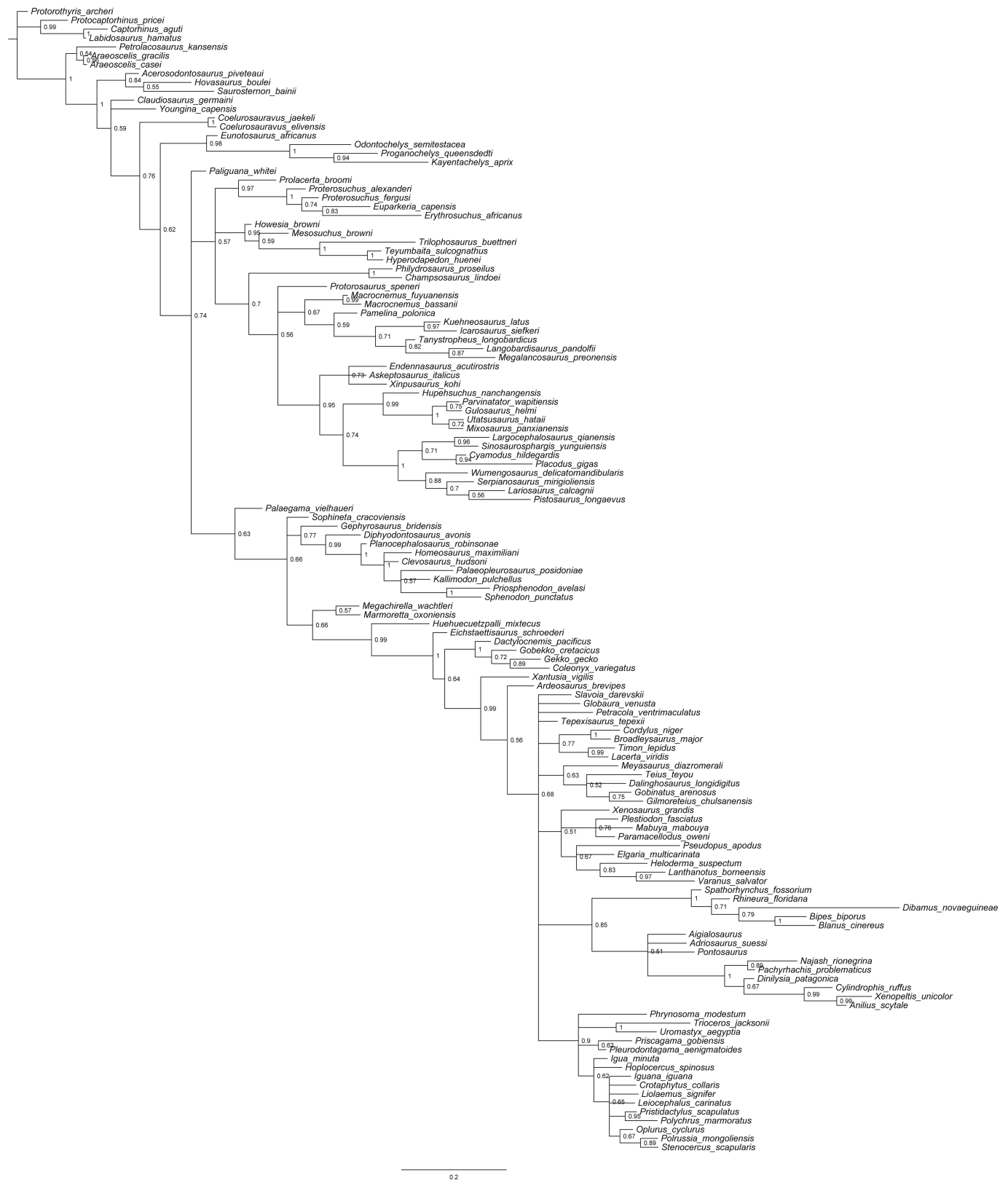
manus in dorsal view. **l**, Line drawing of right manus in dorsal view. Ax.R., axis rib; Ce.Pl., cervical pleurocentrum; Co, cotyle; C.V.3, third cervical vertebra; dc2–5, distal carpals 2–5; DPC, deltopectoral crest; D.R., dorsal rib; D.T., dentary teeth; Epi.St., epiphysal suture; H.Epi., humeral epiphysis; i, intermedium; lc, lateral centrale; Mcl–V, metacarpals I–V; N.A., neural arch; Olf.Tr., olfactory tract; Po.Co., posterior cotyle; Qj.Fr., quadratojugal foramen; Qj.St., quadratojugal suture; r, radiale; Sbd.Sh., subdentary shelf; Sof.Pr., subolfactory processes; u, ulnare. Scale bars, 1 mm (**a**, **b**), 5 mm (**c**, **e**–**h**, **j**–**l**), 10 mm (**d**, **i**).



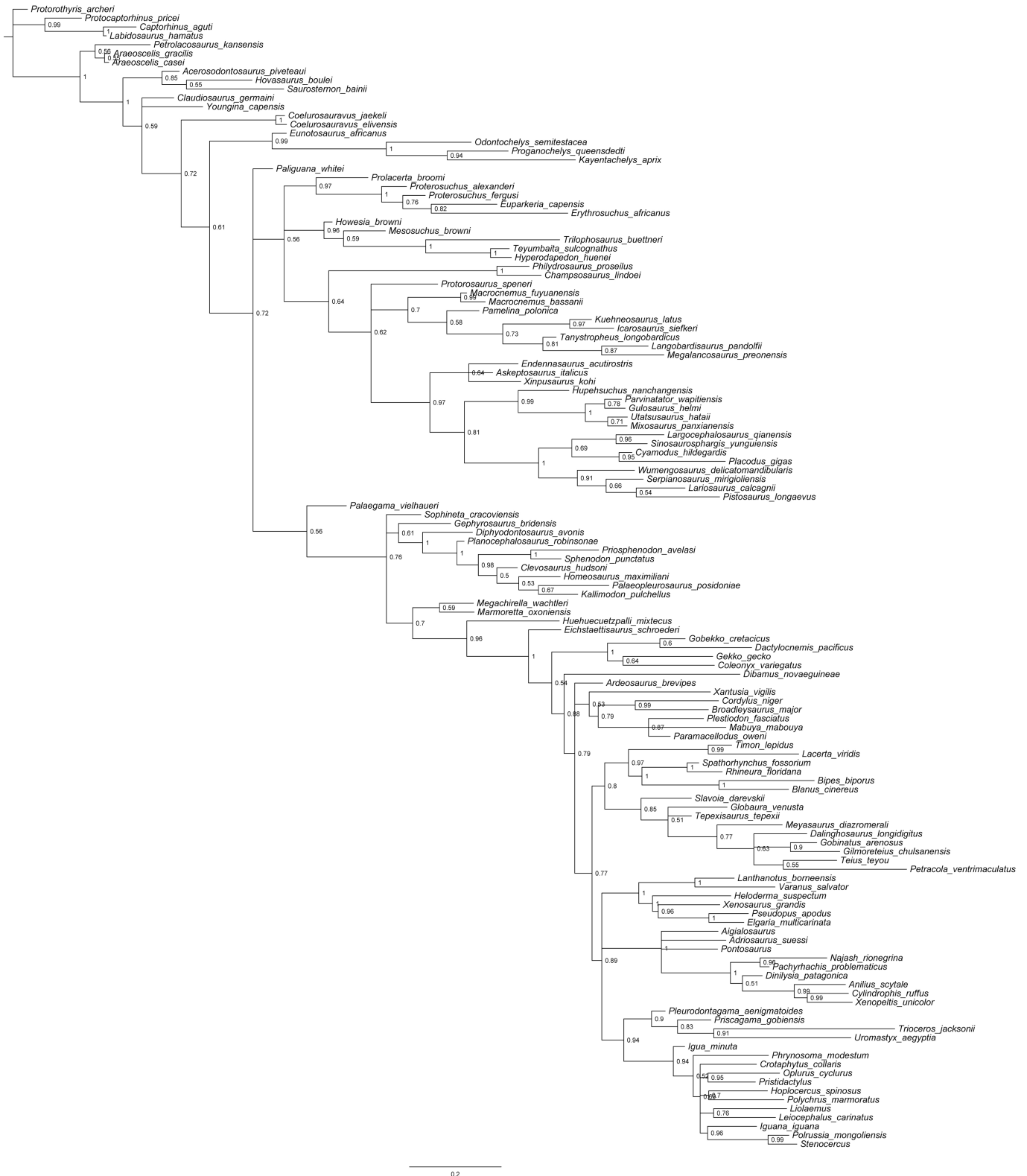
Extended Data Fig. 3 | Equal weights maximum parsimony analysis, morphological data only. Strict consensus of 621 most parsimonious trees (2,268 steps each). Numbers at nodes indicate Bremer indices.



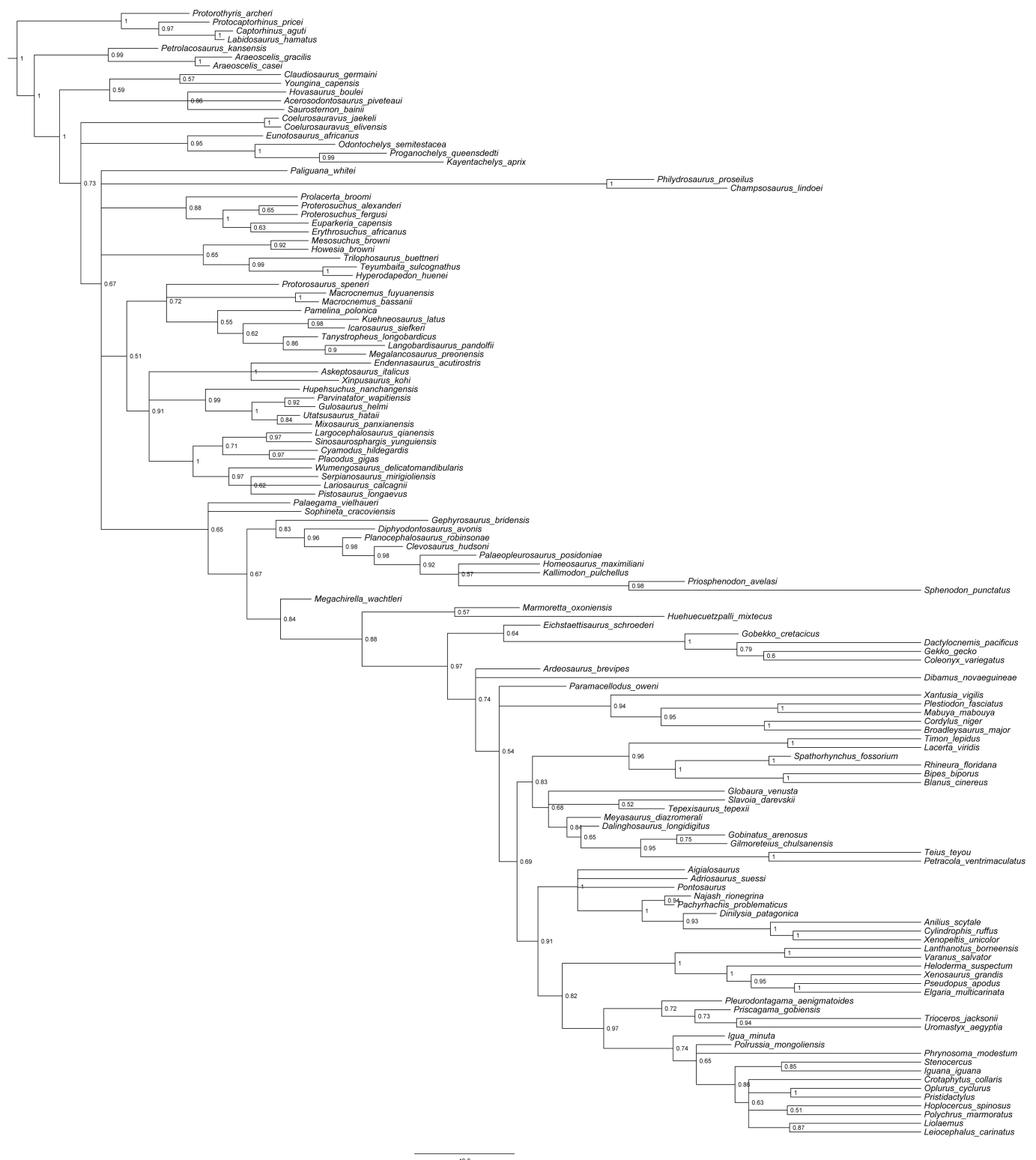
Extended Data Fig. 4 | Implied weighting maximum parsimony analysis, morphological data only. Strict consensus of the five best trees (fit = 91.768892).



Extended Data Fig. 5 | Bayesian inference analysis, morphological data only. Bayesian majority-rule consensus tree. Numbers at nodes indicate posterior probabilities.

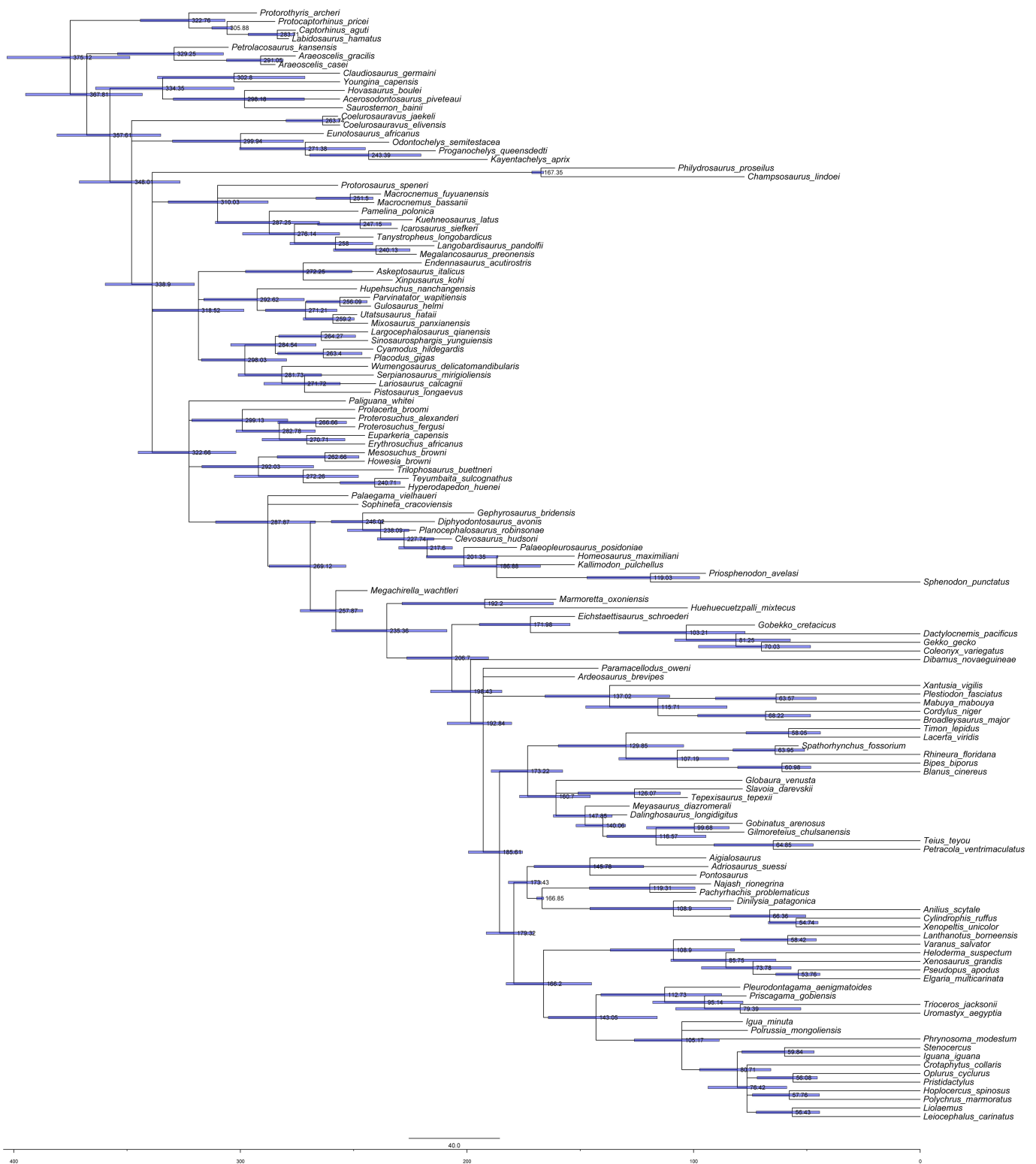


Extended Data Fig. 6 | Bayesian inference analysis, combined morphological and molecular data. Bayesian majority-rule consensus tree. Numbers at nodes indicate posterior probabilities.



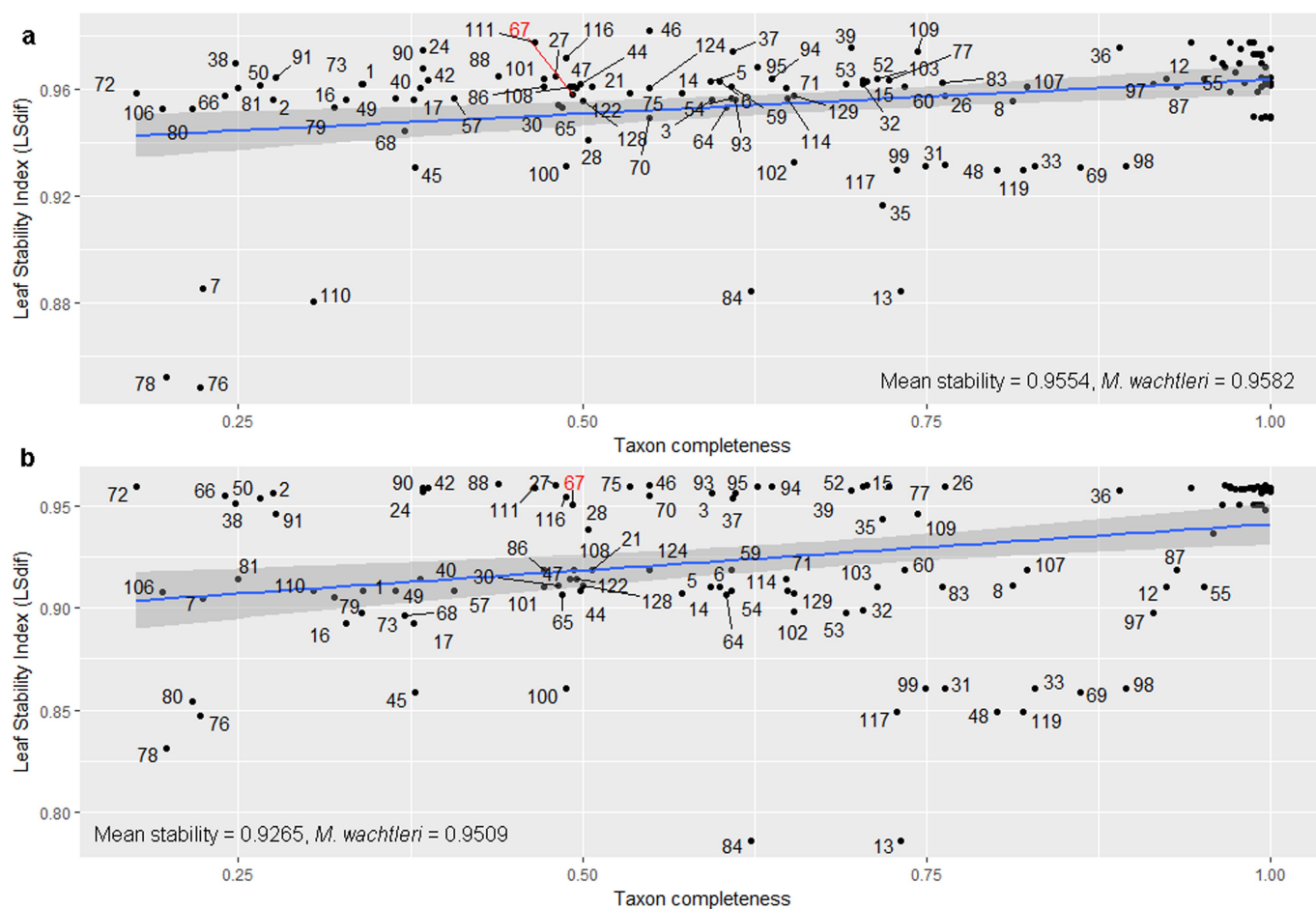
Extended Data Fig. 7 | Relaxed-clock Bayesian inference analysis with total-evidence tip dating using the fossilized birth-death tree model,

combined morphological and molecular data. Bayesian majority-rule consensus tree. Numbers at nodes indicate posterior probabilities.



Extended Data Fig. 8 | Relaxed-clock Bayesian inference analysis with total-evidence tip and node dating using the fossilized birth–death tree model, combined morphological and molecular data. Bayesian

majority-rule consensus tree. Numbers at nodes indicate median estimates for the divergence times, and node bars indicate the 95% highest posterior density for divergence times.



Extended Data Fig. 9 | Taxon stability plotted against taxon completeness in the analysis combining both morphological and molecular data. a, Taxon stability in uncalibrated Bayesian inference analysis. **b**, Taxon stability in relaxed-clock Bayesian inference analysis with tip dating. Taxon stability increases directly proportional to taxon completeness. *M. wachtleri* (taxon 67, in red) has a stability slightly

above average for uncalibrated Bayesian inference, and well above average for Bayesian inference with tip dating. All taxa are identified in Supplementary Table 3 ($n = 129$ taxa). Regression line in blue and 95% confidence interval in grey. Labels for extant taxa ($\sim 100\%$ completeness) are omitted for simplicity.

Long-term effects of species loss on community properties across contrasting ecosystems

Paul Kardol^{1*}, Nicolas Fanin^{1,2} & David A. Wardle^{1,3}

Biodiversity loss can heavily affect the functioning of ecosystems, and improving our understanding of how ecosystems respond to biodiversity decline is one of the main challenges in ecology^{1–4}. Several important aspects of the longer-term effects of biodiversity loss on ecosystems remain unresolved, including how these effects depend on environmental context^{5–7}. Here we analyse data from an across-ecosystem biodiversity manipulation experiment that, to our knowledge, represents the world’s longest-running experiment of this type. This experiment has been set up on 30 lake islands in Sweden that vary considerably in productivity and soil fertility owing to differences in fire history^{8,9}. We tested the effects of environmental context on how plant species loss affected two fundamental community attributes—plant community biomass and temporal variability—over 20 years. In contrast to findings from artificially assembled communities^{10–12}, we found that the effects of species loss on community biomass decreased over time; this decrease was strongest on the least productive and least fertile islands. Species loss generally also increased temporal variability, and these effects were greatest on the most productive and most fertile islands. Our findings highlight that the ecosystem-level consequences of biodiversity loss are not constant across ecosystems and that understanding and forecasting these consequences necessitates taking into account the overarching role of environmental context.

Biodiversity experiments have previously shown that diverse communities are more efficient in capturing resources, and therefore produce more biomass than species-poor communities^{1,2}. Some experiments have also shown that the effects of plant diversity on biomass production increase over time as complementarity of resource use among species increases^{10–12}. This has led to suggestions that ecosystem-level consequences of biodiversity loss might be stronger than predicted from short-term experiments. However, the available evidence emerges from artificially and randomly assembled communities^{10,12}. How the effects of species loss develop over time in natural ecosystems and, importantly, how the long-term effects of biodiversity loss vary among ecosystems remain untested, despite growing evidence that the strength of relationships between plant diversity and productivity can vary with environmental conditions—notably soil resource availability^{6,13}. If soil resource availability is an important driver of plant diversity–productivity relationships, then variation in soil resources could have important consequences for how effects of species loss change over time and in the long term.

The diversity of plant communities can also buffer temporal variability in response to external perturbations and fluctuations in environmental conditions^{14,15}. Greater temporal invariability (the consistency of biomass production over time or $1/(\text{coefficient of variation (CV)})$; also referred to as ‘temporal stability’^{14–17}) in more diverse communities is commonly ascribed to a greater temporal complementarity between species that results from a higher asynchrony of species responses to environmental fluctuations^{14,18}. However, uncertainty remains as to whether there are generalizable relationships between plant diversity

and temporal invariability in ecosystem functioning^{16,19,20}. Although recent evidence suggests that the addition of nitrogen can moderate the effects of species diversity on temporal invariability of community biomass²¹, empirical tests of how these relationships vary with environmental context in natural ecosystems are scarce⁷.

We examined the effects of environmental context on the effects of species loss over time, using data from an across-ecosystem biodiversity manipulation experiment set up in 1996 in a post-fire chronosequence that consists of 30 forested lake islands in northern Sweden^{9,22}. Here, the main disturbance is wildfire: large islands burn more often than smaller ones, which creates a successional gradient across the islands. Large islands have greater soil fertility and greater supply rates of available soil nutrients, and are more productive relative to smaller ones²² (Extended Data Table 1). In line with previous work on this system^{8,9,22}, we divided the islands into three size classes (large (>1.0 ha), medium (0.1–1.0 ha) and small (<0.1 ha)) of ten islands each⁸. For each island, plots were established comprising a full-factorial combination of three dwarf shrub species removals (removal of *Vaccinium myrtillus*, *Vaccinium vitis-idaea* and *Empetrum hermaphroditum*) (see Methods). These species account for >98% of vascular plant biomass in the understorey layer⁹. The ecosystems we studied have comparatively low plant alpha-diversity, as is characteristic of extensive areas of the boreal zone worldwide, and are therefore likely to be especially vulnerable to species loss. Our study design enables us to explore effects of all three-way combinations of the same species across widely contrasting environments.

For each plot we took measurements of the biomass of each shrub species annually from 1996 to 2016 (see Methods); as expected, this revealed generally lower biomass in single- and two-species plots than in three-species plots. However, the effects of removal of certain species, or combinations of species, on total plant biomass strongly differed among island size classes (Fig. 1 and Extended Data Table 1). We predicted that the effects of species loss would be larger on smaller islands, owing to lower soil resource availability and greater resource partitioning among species²³. This would constrain the extent to which the remaining species could compensate for lost species. However, our data instead show that the effects of species loss across different-sized islands are strongly species-dependent (Fig. 1). The biomass of individual species was also strongly affected by plant species removal, with these effects also often varying with island size (Extended Data Fig. 1 and Extended Data Table 1). These species-specific effects are partly explicable by different dominance patterns of the three shrub species among island size classes⁹, which are in turn associated with interspecific differences in nutrient acquisition strategies²⁴.

We then tested whether the strength of the effects of species loss on community biomass increases over time^{10,12}, and if the magnitude of this increase is greater on small islands, as is expected if plants on small—and therefore unproductive—lands were to suffer most from a temporal decrease in resource complementarity. Artificially and randomly assembled biodiversity experiments have shown that complementarity increases over time; for example, through increasing nitrogen

¹Department of Forest Ecology and Management, Swedish University of Agricultural Sciences, Umeå, Sweden. ²INRA, UMR 1391 ISPA, Bordeaux Sciences Agro, Villenave-d’Ornon, France.

³Asian School of the Environment, Nanyang Technological University, Singapore, Singapore. *e-mail: Paul.Kardol@slu.se

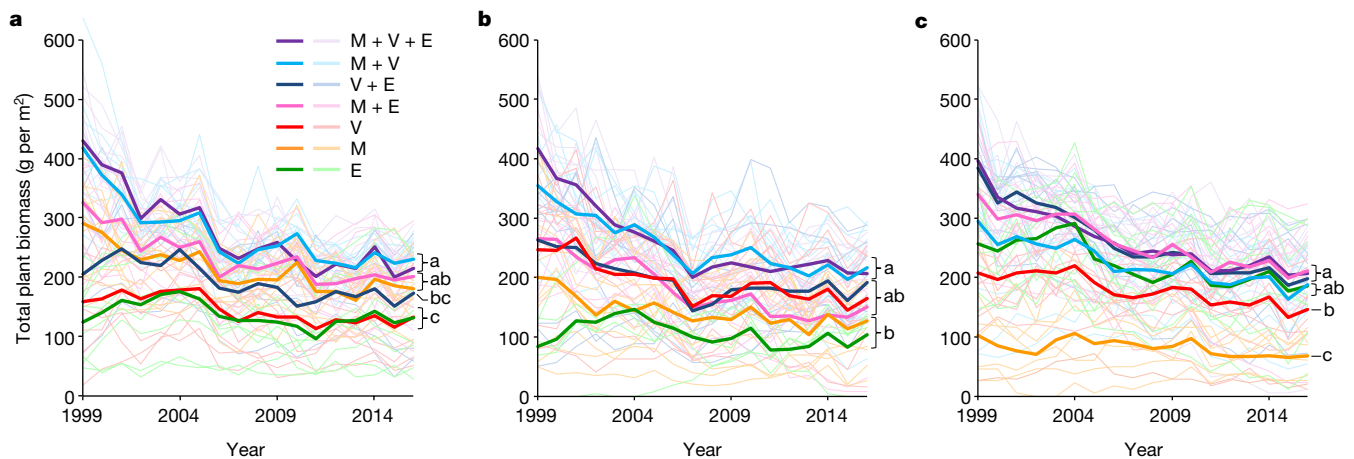


Fig. 1 | Effects of plant species removal on temporal plant biomass patterns. **a–c**, Temporal patterns (1999–2016; years 3–20) of total plant biomass (g per m²) for large (**a**), medium (**b**) and small (**c**) islands are shown. Species codes refer to the plant species remaining (not removed): M, *V. myrtillus*; V, *V. vitis-idaea*; E, *E. hermaphroditum*. Thick dark-coloured lines show mean values per treatment ($n = 10$ islands per size class except for E treatments on large islands ($n = 8$), E treatments on medium islands ($n = 5$), M + E treatments on medium islands ($n = 8$),

E treatments on small islands ($n = 9$), M treatments on small islands ($n = 8$) and M + E treatments on small islands ($n = 9$)). Thin light-coloured lines show values for individual plots. Within island size classes, removal treatments with the same letters are not significantly different across the duration of the study. Treatment effects were tested using linear mixed models fitted by a restricted maximum likelihood method, and we used contrast analyses to test across-year differences between removal treatments (see Methods for details).

retention in more diverse communities²⁵. We found that the effects of species removal on total plant biomass varied through time (Fig. 1 and Extended Data Table 1), but against expectations the strength of the effects of species loss on plant biomass actually decreased. The decrease over time in the amount of variation explained by species richness did not depend on island size (Fig. 2 and Extended Data Table 2). However, the amount of variation explained by removal treatments also decreased, and this was strongest for small islands (Extended Data Fig. 2 and Extended Data Table 2). Polynomial regressions revealed some nonlinearity in these temporal patterns (see Supplementary Information). To further examine the extent to which the effects of species loss depended on context, we also tested how commonly used quantitative measures of the effect of biodiversity on plant biomass

changed over time, and how these temporal changes varied with island size. These analyses indicated that the temporal decrease in the effects of species loss could largely be explained by changes in species complementarity (see Supplementary Information).

Contrary to results obtained from highly controlled, randomly assembled communities^{10,12}, we show with long-term data derived from manipulations of low-diversity, naturally-assembled communities that although biodiversity loss does significantly reduce plant biomass (Fig. 1 and Extended Data Table 1), this effect does not necessarily increase over time. This could occur through the mitigation of the effects of species loss by compensatory responses of remaining species²⁶, with this mitigation strengthening over time. Our findings highlight that such compensatory responses depend greatly on interactions between species identity and environmental context. This indicates that ecosystems are resilient to species loss if other species are able to occupy the newly available niche space. The shape of the resilience pattern of the boreal ecosystem we studied would be determined by the growth responses of the remaining species, owing to the low diversity of this ecosystem. In other systems, colonization rates of species from external pools may be equally or even more important. One possible reason why previous experiments using controlled, randomly assembled communities have often found that the effects of biodiversity increase over time is that they constrain the niche widths of species and do not allow colonization by species from external pools.

Finally, we tested whether species loss increases the temporal variance (or reduces $1/CV$, invariability) in community biomass and whether this increase is strongest for small islands, on the basis that in low-productivity systems the species remaining after biodiversity loss would be less capable of maintaining complementary interactions under temporally fluctuating conditions²⁷. Asynchronous responses among species are particularly important in maintaining biomass production over time, especially in low-productivity systems²¹. In partial support of this, the temporal variability of total plant community biomass differed significantly among species removal treatments and these effects differed with island size, for the most-recent ten years of the experiment (Fig. 3a–c and Extended Data Table 3). Contrary to our expectation, we found that—relative to the three-species mixtures (plots with no species removals)—temporal invariability ($1/CV$) was significantly less in at least three reduced-diversity treatments on medium and large islands, and only one one-species treatment on small islands. Across all removals, we also found significant

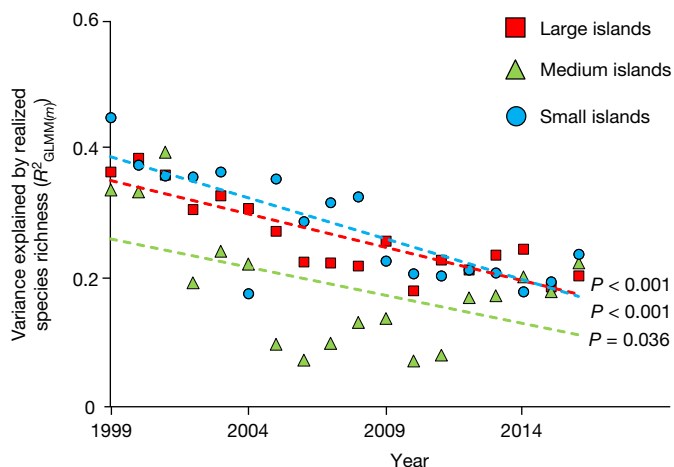


Fig. 2 | Effects of species richness on plant biomass decreases over time. Temporal patterns (1999–2016; years 3–20) of the proportion of variance explained in total plant biomass by realized species richness, for large, medium and small islands. The proportion of variance explained (also called the effect size) was calculated using marginal R^2 values ($R^2_{GLMM(m)}$) for linear mixed models ($n = 10$ islands per size class). Effects of island size, year and their interactions were tested using linear mixed models, and contrasts were applied to compare slopes among island size classes for temporal changes in the amount of variance explained (See Extended Data Table 2 for details). Linear regressions were fitted for each island size class (dotted lines). P values denote slopes significantly different from zero.

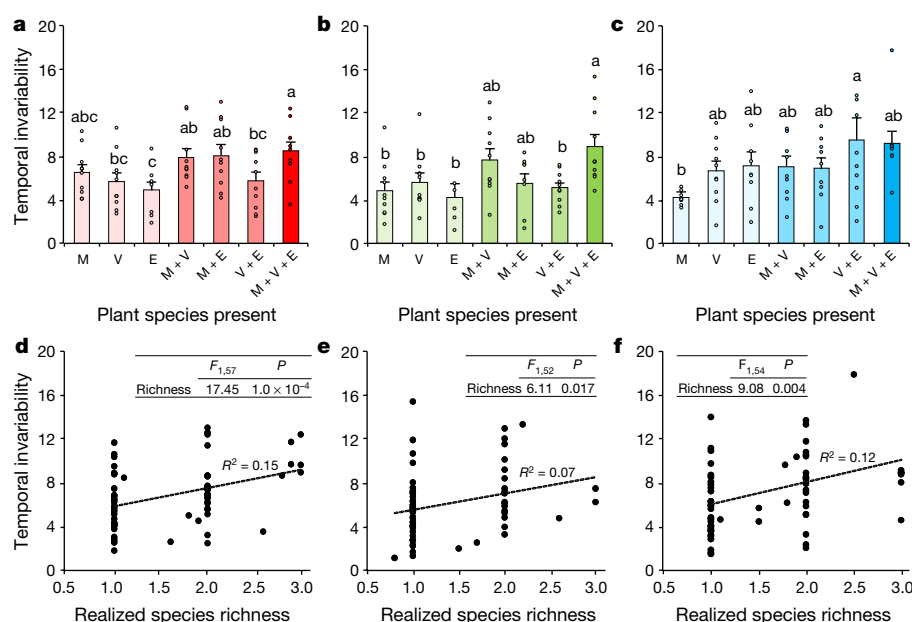


Fig. 3 | Effects of species removal and species richness on temporal biomass invariability. **a–c**, Effects of plant species removals on temporal invariability (1/CV) in community biomass (2007–2016; years 11–20) for large (**a**), medium (**b**) and small (**c**) islands. Species codes (M, E, V) refer to the plant species remaining (not removed). Bar graphs show mean values \pm s.e. ($n = 10$ islands per size class, except for E treatments on large islands ($n = 8$), E treatments on medium islands ($n = 5$), M + E treatments on medium islands ($n = 8$), E treatments on small islands ($n = 9$), M treatments on small islands ($n = 8$), and M + E treatments on small islands ($n = 9$)). Dots indicate values for individual islands. Data were analysed

using linear mixed models (see Methods for details). Within island size classes, removal treatments with the same letters are not significantly different across years (Tukey's post hoc test, $P < 0.05$). **d–f**, Relationship between mean realized species richness and temporal invariability (2007–2016) for large (**d**), medium (**e**) and small (**f**) islands. Dots indicate individual plots (large islands, $n = 68$; medium islands, $n = 63$; small islands, $n = 65$). Insets show results from linear models testing for the effects of species richness on temporal invariability (see Methods for model details). Significant P values (< 0.05) are shown. R^2 values indicate the strength of the relationship.

positive relationships between realized species richness and temporal invariability in community biomass that were independent of island size (Fig. 3d–f and Extended Data Table 3). When analysed for the initial years of our experiment, in which there was an overall downward trend in biomass (see Methods), or for the full experimental period, the effects of species removal on temporal variability were less pronounced (Supplementary Table 8 and Supplementary Figs. 11, 12). However, our findings support theoretical work²⁸ that proposes that the negative effects of species loss on the consistency of biomass production over time depend strongly on which species are lost and the compensatory dynamics of remaining species over time, and highlight that these effects are mediated by environmental context and associated differences in plant dominance.

Biodiversity loss is known to have important consequences for the structure and productivity of plant communities and associated ecosystem functions^{1–4}. Most available empirical evidence is from studies conducted under highly controlled conditions in which plant diversity has been varied through random draws from species pools, but our results show that different patterns may occur when species are removed from natural ecosystems^{10,12}. Moreover, we show that the long-term effects of diversity loss are not consistent and can differ greatly across contrasting ecosystems. Although human impacts often cause the loss of biodiversity across contrasting types of ecosystems²⁹, our results reveal that this loss of biodiversity affects the functioning of different ecosystems in contrasting ways. The loss of biodiversity in natural ecosystems under global change is not a random process, and we show that—for a type of low-diversity system that is globally widespread but little studied from a diversity-functioning perspective—it matters not only which species are lost and to what extent the remaining species are able to exploit the available resources, but also from which ecosystems they are lost. In acknowledging the need for testing the generality of our findings for other systems, including more-diverse systems that have been much more widely studied, we emphasize that forecasting the consequences of biodiversity loss for the functioning

of natural ecosystems necessitates taking into account when and how environmental context moderates the effects of this loss.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0138-7>.

Received: 27 February 2017; Accepted: 16 April 2018;

Published online: 23 May 2018

- Hector, A. et al. Plant diversity and productivity experiments in European grasslands. *Science* **286**, 1123–1127 (1999).
- Hooper, D. U. et al. A global synthesis reveals biodiversity loss as a major driver of ecosystem change. *Nature* **486**, 105–108 (2012).
- Isbell, F. et al. Linking the influence and dependence of people on biodiversity across scales. *Nature* **546**, 65–72 (2017).
- Duffy, J. E., Godwin, C. M. & Cardinale, B. J. Biodiversity effects in the wild are common and as strong as key drivers of productivity. *Nature* **549**, 261–264 (2017).
- Craven, D. et al. Plant diversity effects on grassland productivity are robust to both nutrient enrichment and drought. *Phil. Trans. R. Soc. Lond. B* **371**, 20150277 (2016).
- Fridley, J. D. Resource availability dominates and alters the relationship between species diversity and ecosystem productivity in experimental plant communities. *Oecologia* **132**, 271–277 (2002).
- Gross, K. et al. Species richness and the temporal stability of biomass production: a new analysis of recent biodiversity experiments. *Am. Nat.* **183**, 1–12 (2014).
- Wardle, D. A., Hörnberg, G., Zackrisson, O., Kalela-Brundin, M. & Coomes, D. A. Long-term effects of wildfire on ecosystem properties across an island area gradient. *Science* **300**, 972–975 (2003).
- Wardle, D. A. & Zackrisson, O. Effects of species and functional group loss on island ecosystem properties. *Nature* **435**, 806–810 (2005).
- Cardinale, B. J. et al. Impacts of plant diversity on biomass production increase through time because of species complementarity. *Proc. Natl Acad. Sci. USA* **104**, 18123–18128 (2007).
- Guerrero-Ramírez, N. R. et al. Diversity-dependent temporal divergence of ecosystem functioning in experimental ecosystems. *Nat. Ecol. Evol.* **1**, 1639–1642 (2017).

12. Reich, P. B. et al. Impacts of biodiversity loss escalate through time as redundancy fades. *Science* **336**, 589–592 (2012).
13. Reich, P. B. et al. Plant diversity enhances ecosystem responses to elevated CO₂ and nitrogen deposition. *Nature* **410**, 809–812 (2001).
14. Isbell, F. I., Polley, H. W. & Wilsey, B. J. Biodiversity, productivity and the temporal stability of productivity: patterns and processes. *Ecol. Lett.* **12**, 443–451 (2009).
15. Tilman, D., Reich, P. B. & Knops, J. M. H. Biodiversity and ecosystem stability in a decade-long grassland experiment. *Nature* **441**, 629–632 (2006).
16. Bezemer, T. M. & van der Putten, W. H. Diversity and stability in plant communities. *Nature* **446**, E6–E7 (2007).
17. Cardinale, B. J. et al. Biodiversity simultaneously enhances the production and stability of community biomass, but the effects are independent. *Ecology* **94**, 1697–1707 (2013).
18. Morin, X., Fahse, L., de Mazancourt, C., Scherer-Lorenzen, M. & Bugmann, H. Temporal stability in forest productivity increases with tree diversity due to asynchrony in species dynamics. *Ecol. Lett.* **17**, 1526–1535 (2014).
19. van Ruijven, J. & Berendse, F. Contrasting effects of diversity on the temporal stability of plant populations. *Oikos* **116**, 1323–1330 (2007).
20. de Mazancourt, C. et al. Predicting ecosystem stability from community composition and biodiversity. *Ecol. Lett.* **16**, 617–625 (2013).
21. Hautier, Y. et al. Eutrophication weakens stabilizing effects of diversity in natural grasslands. *Nature* **508**, 521–525 (2014).
22. Wardle, D. A. et al. Linking vegetation change, carbon sequestration and biodiversity: insights from island ecosystems in a long-term natural experiment. *J. Ecol.* **100**, 16–30 (2012).
23. Huston, M. A. & DeAngelis, D. L. Competition and coexistence: the effects of resource transport and supply rates. *Am. Nat.* **144**, 954–977 (1994).
24. Gundale, M. J., Hyodo, F., Nilsson, M. C. & Wardle, D. A. Nitrogen niches revealed through species and functional group removal in a boreal shrub community. *Ecology* **93**, 1695–1706 (2012).
25. Fargione, J. et al. From selection to complementarity: shifts in the causes of biodiversity–productivity relationships in a long-term biodiversity experiment. *Proc. R. Soc. Lond. B* **274**, 871–876 (2007).
26. Symstad, A. J. & Tilman, D. Diversity loss, recruitment limitation, and ecosystem functioning: lessons learned from a removal experiment. *Oikos* **92**, 424–435 (2001).
27. Loreau, M. Biodiversity and ecosystem functioning: recent theoretical advances. *Oikos* **91**, 3–17 (2000).
28. Fowler, M. S. et al. Species dynamics alter community diversity–biomass stability relationships. *Ecol. Lett.* **15**, 1387–1396 (2012).
29. Suding, K. N. et al. Functional- and abundance-based mechanisms explain diversity loss due to N fertilization. *Proc. Natl Acad. Sci. USA* **102**, 4387–4392 (2005).

Acknowledgements We thank numerous assistants for help in the field. This work was supported by grants to D.A.W. from the Swedish Research Council (Vetenskapsrådet) and a Wallenberg Scholars award.

Reviewer information Nature thanks Y. Hautier, P. Morin and F. van der Plas for their contribution to the peer review of this work.

Author contributions D.A.W. acquired the necessary funding, designed the experiment and collected the data. N.F. and P.K. analysed the data in close consultation with D.A.W. P.K. wrote the first draft of the manuscript, and all authors contributed to manuscript completion and revision.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0138-7>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0138-7>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to P.K.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. Sample size (that is, the numbers of islands used) was based on analysis of data from a previous study in this system using a larger number of islands³⁰. The experimental treatments were randomized, but investigators were not blinded to treatments during measurement because it would not have been tractable to do so.

Study system. This study was conducted along a post-fire chronosequence consisting of 30 forested islands in lakes Hornavan and Uddjaure in northern Sweden (65° 55' N to 66° 09' N; 17° 43' E to 17° 55' E). The islands are all of the same geologic age and experience the same macroclimate. Mean annual precipitation is 750 mm, of which 350 mm falls during the growing season (May to September), and the mean annual temperature is 0.5 °C (13 °C in July and −14 °C in January). The only major extrinsic factor that varies among the islands is the frequency of wildfire, with larger islands burning more frequently than smaller islands because they have a greater probability of being struck by lightning^{8,30}. Previous work has shown that as islands become smaller and the time since fire increases, the availability of nutrients (notably nitrogen and phosphorus) decreases; this leads to impairment of the rates of decomposition and nutrient fluxes, and lower plant biomass and productivity^{8,22,30–32}. Consistent with previous studies in this system^{8,22,30–32}, these islands were divided into three size classes with 10 islands per class: large (>1.0 ha), medium (0.1–1.0 ha) and small (<0.1 ha), with a mean time since last major fire of about 585, 2,180 and 3,250 years at the start of the experiment, respectively⁸. This replicate number was necessary for correctly identifying important differences among islands size classes³⁰. The overstory vegetation is dominated by *Betula pubescens*, *Pinus sylvestris* and *Picea abies*, and the ground-layer vegetation consists of the ericaceous dwarf shrubs *V. myrtillus*, *V. vitis-idaea*, and *E. hermaphroditum*, and feather mosses. Extended Data Table 4 shows a selection of ecosystem properties for each of the three island size classes.

Experimental design. For this study we focused on the dwarf shrub vegetation, and used a removal experiment approach because this is a powerful tool for investigating the effects of local, non-random losses of biotic components and species interactions in natural ecosystems³³. We established 8 experimental plots on each of the 30 islands, each representing a different dwarf shrub species removal treatment; that is, full factorial removal of *V. myrtillus*, *V. vitis-idaea*, and *E. hermaphroditum*⁹. One of these plots involved removal of all shrub species and was not considered in this study. The level of species richness we manipulated is typical of the diversity of understory vegetation that characterizes boreal forests at the spatial scale of our plots. As such, our system is representative of low-diversity communities for which ecosystem functioning is especially vulnerable to species loss. *V. myrtillus*, *V. vitis-idaea* and *E. hermaphroditum* dominate the ericaceous shrub layer in large, medium and small islands, respectively, and collectively account for ± 98% of vascular plant biomass in the understory⁸ (see Supplementary Table 9 for details on productivity and leaf traits for each of the shrub species across the island size gradient). All plots were 55 × 55 cm, but only the inner 45 × 45 cm was measured. All plots were located at similar distances from the shore for each island, regardless of island size, to prevent edge and microclimatic effects from confounding the results^{8,9}. The experiment was established in August 1996 and has been maintained annually ever since. Shrub removal treatments were conducted and maintained through annual physical removal of vegetation. We recognize that vegetation removals impose initial disturbance effects, but these are likely to be transient and of minimal importance after the initial years^{34,35}. In a separate 14-year experiment performed on the dwarf shrub community on all 30 islands and in plots that are adjacent to the plots used in this study, and that involved experimental disturbance treatments with greater disturbance than the removals performed in this study³⁶, it was shown that disturbance legacies of vegetation removal are mostly gone within 3–6 years. Further details of the removal experiment reported here have previously been presented⁹.

Every August from 1996 until 2016 inclusive, the total cover of shrub species was assessed in each plot by point quadrat analysis, by determining the total number of times the vegetation of that species was intercepted by a total of 100 downwardly projecting points. The total number of intercepts for each species was then converted to biomass per unit area through equations previously developed by destructively sampling calibration plots⁸. For each of the three shrub species the total number of point intercepts is very closely correlated with aboveground standing biomass, with R^2 values that are consistently above 0.90⁸. During the first ten years of the experiment, we observed an overall downward trend in plant community biomass across island sizes. This trend was not related to the removal treatments; it also occurred in the control plots (Fig. 1). There are two plausible explanations. First, it could have been a function of where we positioned the plots when the experiment was set up in 1996, as we always placed our plots in well-developed vegetation. As part of their natural population dynamics, long-lived dwarf shrub species in boreal forest tend to move around over time³⁷. This means that well-developed shrub patches may well decline in biomass over time whereas less developed shrub patches may well aggrade to become well-developed patches

in the future. Second, over the course of the experiment plant biomass was likely to have turned over only about three times^{9,38}, and if during one of the cycles there is high (or low) biomass, then this high (or low) biomass situation will persist for some years. It is plausible that if there were environmental conditions that promoted high biomass in all plots at the start of the experiment then that effect would have a legacy that would persist several years into the experiment. We emphasize that this downward trend is a non-biased pattern that has no bearing on conclusions that we can draw from this study.

Data analyses. To account for the initial disturbance effects of the removal treatments⁹, the first three years of data (1996–1999) were excluded from the data analyses. Further, in all analyses we used only plots in which at least one of the three shrub species was present at the start of the experiment (that is, after implementation of the removal treatments). This resulted in the a priori exclusion of 13 out of a total of 210 plots across all islands (large islands: 2 plots from which *V. myrtillus* and *V. vitis-idaea* were removed; medium islands: 5 plots from which *V. myrtillus* and *V. vitis-idaea* were removed and two plots where *V. vitis-idaea* was removed; small islands: 1 plot from which *V. myrtillus* and *V. vitis-idaea* were removed, 1 plot from which *V. vitis-idaea* was removed, and two plots from which *V. vitis-idaea* and *E. hermaphroditum* were removed).

Effects of plant species removal, island size class and their interaction on total plant biomass and biomass of each of the three shrub species were tested using linear mixed models (LMMs) fitted by a restricted maximum likelihood method. Island identity was included as a random effect to account for the repeated measurements across time and plots were nested within island identity to enable comparison of treatments within each island separately. We fitted a first-order auto-regressive variance structure¹¹. To further explore the effects of species removal on total and plant- species-specific biomass, we ran LMMs separately for each island size class and used contrast analyses to test across-year differences between removal treatments. To evaluate potential effects of the overall downward trend in plant biomass during the first ten years of the experiment, we ran all LMMs separately for 1999–2006, 2007–2016 and 1999–2016. These separate analyses showed that effects of species removal and island size were largely consistent across these different stages of the experiment (see Extended Data Table 1).

To test how the effects of species loss on plant community biomass changed over time, we calculated the proportion of variance explained by the species removal treatment and by species richness (to better account for species lost over the course of the experiment) separately for each year of the experiment (1999–2016) and for each island size class (small, medium and large). For each plot, the proportion of variance explained (also called effect size) was calculated using marginal R^2 values ($R^2_{\text{GLMM}(m)}$) for LMMs³⁹, with species removals or realized species richness (that is, the number of shrub species present across the entire experimental period³⁵) considered as fixed factors and island identity as a random factor. The values of the marginal $R^2_{\text{GLMM}(m)}$ describe the proportion of variance explained by fixed factors alone. We then tested how $R^2_{\text{GLMM}(m)}$ changed over time using LMMs with island size class and year as fixed factors and year as a continuous variable. Contrasts on the interaction between island size class and year were used to test whether the slopes of regressions between year and $R^2_{\text{GLMM}(m)}$ differed between island size classes. Finally, we used polynomial models to test for nonlinear patterns. In these models, we employed LMMs with island size class and the second-order polynomial of year as fixed factors and year as a continuous variable. To compare whether polynomial models varied among island size classes, we compared models by grouping island size classes two-by-two and tested whether the effect of the grouping variable was significant.

To better understand the underlying mechanisms of the effects of species loss on total and plant-species-specific biomass and how they change over time, we used two approaches for comparing biomass in multi-species communities relative to their component monocultures⁴⁰; in our study, monocultures consisted of the three treatments in which only one of the three species was not removed. These measures were calculated separately for three-species mixtures (that is, plots for which no removals were performed) and for each of the three possible two-species mixtures (that is, plots for which one of the three species was removed). First, we used the additive partitioning approach to separate net biodiversity effects into complementarity and selection effects⁴¹. Net biodiversity effects measure the deviation of biomass in any given multi-species community from its expected value based on biomass of its component species grown in monoculture. Complementarity effects measure the average relative species biomass in multi-species communities relative to the expected biomass based on the weighted-average biomass of their component monocultures. Selection effects measure the effects of species with higher or lower than average biomass in monocultures on total community biomass. Second, we calculated two measures of complementarity that are independent of absolute biomass values: relative yield totals and transgressive overyielding. The relative yield of a species measures its biomass in a multi-species community as a proportion of its biomass in monoculture, and the relative yield total of a multi-species community is the sum of the relative yield of all component species⁴². Transgressive

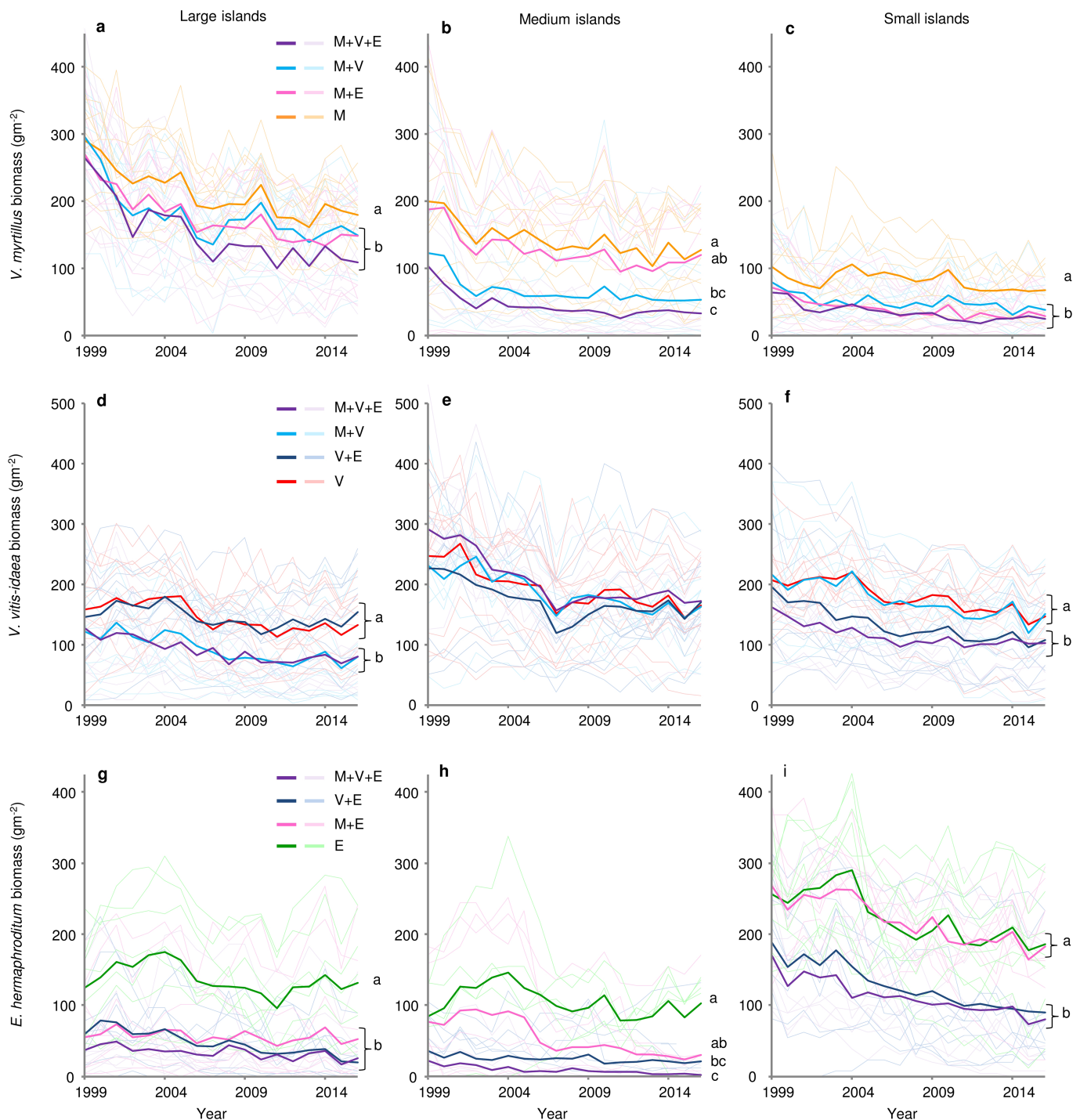
overyielding measures whether multi-species communities obtain higher biomass than the monoculture of its most productive component species. Transgressive overyielding was calculated as D_{\max} following a previously published method⁴³. We then tested how each of the measures of biodiversity effects that we calculated changed over time using LMMs and polynomial models as described above for $R^2_{\text{GLMM}(m)}$, and with individual islands ($n = 10$ for each of the three island size classes) used as units of replication. We fitted a first-order auto-regressive variance structure to account for temporal pseudoreplication through repeated measurements on each island¹¹. Measures of biodiversity effects could not be calculated whenever calculations included denominator values of zero, and there were also some cases (<1%) in which very low denominator biomasses resulted in extremely high values of biodiversity measures; these extreme values were removed from the analyses following the outlier labelling rule with a conservative tuning parameter of $g = 2.2^{44}$.

Finally, to test how species loss affects temporal invariability in plant community biomass, we calculated the inverse coefficient of variation ($1/\text{CV}$); that is, the mean biomass over the study period divided by the standard deviation for that period^{14–17}. The inverse of coefficient of variation provides a widely used standardized measure of invariability that is comparable across ecosystems, and that is commonly referred to as ‘temporal stability’⁷. We focused on temporal invariability for the most-recent ten-year period, that is, 2007–2016. This is consistent with the ten-year period used in comparable analyses of temporal invariability¹⁵, and was done because up to 2007 there was a general decline in total plant community biomass, whereas over the last ten years of the experiment there was no apparent directional change in total community biomass (Fig. 1). However, we also calculated temporal invariability for 1999–2006 and for 1999–2016. For each period, effects of plant species removal and island size class on temporal invariability were then tested using LMMs with island identity as random factor. Within island size classes, contrast analyses were used to test differences between species removal treatments. Finally, relationships between realized species richness and temporal invariability were tested using linear regression. Similar to the biomass data, all analyses for invariability were run separately for 1999–2006, 2007–2016 and 1999–2016.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

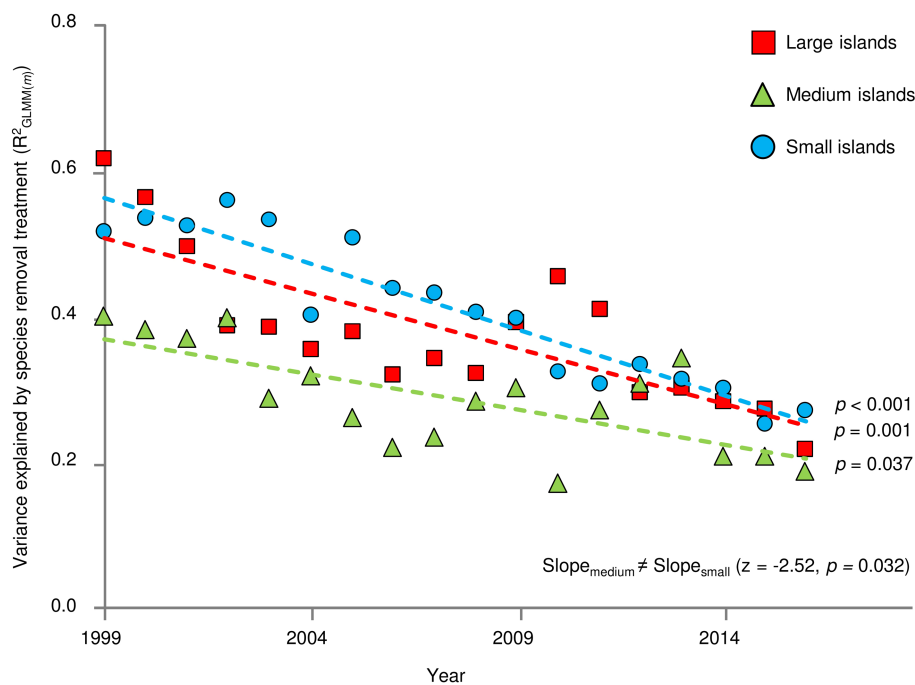
Data availability. The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

30. Wardle, D. A., Zackrisson, O., Hörnberg, G. & Gallet, C. The influence of island area on ecosystem properties. *Science* **277**, 1296–1299 (1997).
31. Wardle, D. A., Walker, L. R. & Bardgett, R. D. Ecosystem properties and forest decline in contrasting long-term chronosequences. *Science* **305**, 509–513 (2004).
32. Clemmensen, K. E. et al. Roots and associated fungi drive long-term carbon sequestration in boreal forest. *Science* **339**, 1615–1618 (2013).
33. Díaz, S., Symstad, A. J., Chapin, F. S., Wardle, D. A. & Huenneke, L. F. Functional diversity revealed by removal experiments. *Trends Ecol. Evol.* **18**, 140–146 (2003).
34. Coomes, D. A. & Grubb, P. J. Impacts of root competition in forests and woodlands: a theoretical framework and review of experiments. *Ecol. Monogr.* **70**, 171–207 (2000).
35. McGrady-Steed, J., Harris, P. M. & Morin, P. J. Biodiversity regulates ecosystem predictability. *Nature* **390**, 162–165 (1997).
36. Wardle, D. A. & Jonsson, M. Long-term resilience of above- and below ground ecosystem components among contrasting ecosystems. *Ecology* **95**, 1836–1849 (2014).
37. Flower-Ellis, J. G. K. *Age Structure and Dynamics in Stands of Bilberry (Vaccinium myrtillus L.)* (Department of Forest Ecology and Forest Soils, Royal College of Forestry, Stockholm, 1971).
38. Wardle, D. A. et al. Drivers of inter-year variability of plant production and decomposers across contrasting island ecosystems. *Ecology* **93**, 521–531 (2012).
39. Nakagawa, S. & Schielzeth, H. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods Ecol. Evol.* **4**, 133–142 (2013).
40. Roscher, C. et al. Overyielding in experimental grassland communities – irrespective of species pool or spatial scale. *Ecol. Lett.* **8**, 419–429 (2005).
41. Loreau, M. & Hector, A. Partitioning selection and complementarity in biodiversity experiments. *Nature* **412**, 72–76 (2001).
42. Hector, A. The effect of diversity on productivity: detecting the role of species complementarity. *Oikos* **82**, 597–599 (1998).
43. Loreau, M. Separating sampling and other effects in biodiversity experiments. *Oikos* **82**, 600–602 (1998).
44. Hoaglin, D. C. & Iglewicz, B. Fine-tuning some resistant rules for outlier labeling. *J. Am. Stat. Assoc.* **82**, 1147–1149 (1987).
45. Lagerström, A., Esberg, C., Wardle, D. A. & Giesler, R. Soil phosphorus and microbial response to a long-term wildfire chronosequence in northern Sweden. *Biogeochemistry* **95**, 199–213 (2009).



Extended Data Fig. 1 | Effects of plant species removal on temporal biomass patterns (1999–2016, years 3–20) of individual species biomass. a–i, Data show individual species biomass (g per m²) for *V. myrtillus* (a–c), *V. vitis-idaea* (d–f) and *E. hermaphroditum* (g–i) for large, medium and small islands. Species codes (M, V, E) refer to the plant species remaining after removal. Thick dark-coloured lines show mean values per treatment ($n = 10$ islands per size class, except for E treatments on large islands ($n = 8$), E treatments on medium islands ($n = 5$), M + E treatments on medium islands ($n = 8$), E treatments on small

islands ($n = 9$), M treatments on small islands ($n = 8$) and M + E treatments on small islands ($n = 9$). Thin light-coloured lines show values for individual plots. Within island size classes, removal treatments with the same letters are not significantly different across years through the duration of the study. Treatment effects were tested using linear mixed models fitted by a restricted maximum likelihood method, and we used contrast analyses to test across-year differences between removal treatments (see Methods for details).



Extended Data Fig. 2 | Temporal patterns (1999–2016; years 3–20) of the proportion of variance in total plant biomass explained by the species-removal treatment for large, medium and small islands. The proportion of variance explained (also called the effect size) was calculated using marginal R^2 values ($R^2_{\text{GLMM}(m)}$) for linear mixed models ($n = 10$ islands per size class). Linear regressions were fit for each island size

class (dotted lines). We used linear mixed models to test how $R^2_{\text{GLMM}(m)}$ changed over time with island size class and year as fixed factors and year as a continuous variable. Contrasts on the interaction between island size class and year were used to test if the slopes of regressions between year and $R^2_{\text{GLMM}(m)}$ differed between island size classes. Significant differences in slopes among island size classes at $\alpha = 0.05$ are indicated in the panel.

Extended Data Table 1 | Effects of species removal, island size and their interactions on total and species-specific plant biomass

Period	Source of variation	Total biomass			<i>V. myrtillus</i>			<i>V. vitis-idaea</i>			<i>E. hermaphroditum</i>		
		df	<i>F</i>	<i>p</i>	df	<i>F</i>	<i>p</i>	df	<i>F</i>	<i>p</i>	df	<i>F</i>	<i>p</i>
1999-2016	Species removal (SR)	6,149	23.24	<0.001	3,75	16.35	<0.001	3,80	3.51	0.019	3,69	22.09	<0.001
	Island size (IS)	2,27	0.33	0.721	2,27	15.60	<0.001	2,27	4.00	0.030	2,27	20.04	<0.001
	SR × IS	12,149	6.10	<0.001	6,75	2.49	0.030	6,80	5.28	<0.001	6,69	1.71	0.133
1999-2006	Species removal	6,149	23.55	<0.001	3,75	10.37	<0.001	3,80	2.85	0.043	3,69	20.93	<0.001
	Island size (IS)	2,27	0.69	0.509	2,27	15.83	<0.001	2,27	4.46	0.021	2,27	19.05	<0.001
	SR × IS	12,149	6.24	<0.001	6,75	1.85	0.102	6,80	3.95	0.002	6,69	1.73	0.128
2007-2016	Species removal	6,149	14.60	<0.001	3,75	26.15	<0.001	3,80	4.19	0.008	3,69	17.54	<0.001
	Island size (IS)	2,27	0.52	0.598	2,27	13.60	<0.001	2,27	3.14	0.060	2,27	19.54	<0.001
	SR × IS	12,149	4.94	<0.001	6,75	3.06	0.001	6,80	5.45	<0.001	6,69	1.41	0.224

Data were analysed using linear mixed models and were analysed for 1999–2016, as well as separately for 1999–2006 and 2007–2016 ($n = 10$ islands per size class) (see Methods for details). Significant *P* values (<0.05) are shown as bold numbers. Results from contrast analyses testing across-year differences in total and plant-species-specific biomass between removal treatments for large, medium and small islands are indicated in Fig. 1 and Extended Data Fig. 1 (1999–2016) and in Supplementary Table 10 (1999–2006 and 2007–2016). For the analyses of total biomass all seven species removal treatments were used; for each of the analyses for *V. myrtillus*, *V. vitis-idaea* and *E. hermaphroditum*, only the treatments from which those species were not removed were used.

Extended Data Table 2 | Effects of species removal, island size and their interactions on the amount of variance explained ($R^2_{\text{GLMM}(m)}$) by species removal treatment and by realized species richness

Source of variation	df	Amount of variance explained by			
		species removal		species richness	
		<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>
Island size	2,48	3.33	0.044	0.54	0.589
Year	1,48	107.0	<0.001	45.3	<0.001
Island size × Year	2,48	3.28	0.046	0.52	0.596
Slope comparisons		<i>z</i>	<i>p</i>	<i>z</i>	<i>p</i>
Large vs. medium		1.66	0.2208	0.39	0.918
Large vs. small		-0.86	0.6667	-0.62	0.808
Medium vs small		-2.52	0.0317	-1.02	0.567

The proportion of variance explained (also called the effect size) was calculated using marginal R^2 values ($R^2_{\text{GLMM}(m)}$) for linear mixed models ($n = 10$ islands per size class). Data were analysed using linear mixed models, and contrasts were applied to compare slopes among island size classes for temporal changes in the amount of variance explained. Significant P values (<0.05) are shown as bold numbers. See Methods for model details.

Extended Data Table 3 | Effects of species removal or realized species richness, and island size, and their interactions on temporal invariability

	Source of variation	df	<i>F</i>	<i>p</i>
A	Species removal (SR)	6,148	6.10	<0.001
	Island size (IS)	2,27	1.23	0.309
	SR × IS	12,148	2.13	0.018
B	Realized species richness (R)	1,163	30.06	<0.001
	Island size (IS)	2,27	0.46	0.637
	R × IS	2,163	0.12	0.885

Data were analysed using linear models testing for the effects of species removal, island size (small, medium and large) and their interactions (A); and realized species richness, island size and their interactions on temporal invariability ($1/(\text{coefficient of variation or CV})$) in plant community biomass (B). Data used are for the period 2007–2016 ($n = 10$ islands per size class). Significant *P* values (<0.05) are shown as bold numbers.

Extended Data Table 4 | Selected ecosystem properties across the island size gradient

Ecosystem property	Island size		
	Small	Medium	Large
Time since last fire (years)	3250 ± 439 a	2180 ± 385 b	585 ± 233 c
Net primary productivity (g/m ² /yr)	159 ± 18 b	247 ± 12 a	256 ± 14 a
Standing plant biomass (g/m ²)	3470 ± 470 b	8340 ± 877 a	9349 ± 485 a
Vascular plant species richness (number of species in a 10 m radius circular plot)	10.6 ± 0.6 a	8.6 ± 0.4 b	6.6 ± 0.6 c
Humus C to N ratio	32.9 ± 0.79 b	36.0 ± 1.17 ab	40.4 ± 1.18 a
Humus C to P ratio	759 ± 30 a	687 ± 36 ab	623 ± 20 b
Humus N to P ratio	23.3 ± 1.1 a	19.1 ± 0.9 b	15.4 ± 0.5 c
Mineral N (MIN) (µgN/g)	25.3 ± 8.0 b	58.1 ± 9.2 a	38.2 ± 14.4 ab
Dissolved organic N (DON) (µgN/g)	40.3 ± 4.6 b	50.7 ± 5.5 a	39.1 ± 7.2 b
MIN/(MIN+DON)	0.39 ± 0.03 b	0.53 ± 0.05 a	0.49 ± 0.04 a
Mineral P (µgP/g)	24.4 ± 2.3 b	37.7 ± 4.3 a	43.6 ± 4.9 a
Membrane-extractable P (mmol/kg)	4.9 ± 0.3 b	6.5 ± 0.4 a	5.9 ± 0.7 ab
Light transmission (%)	68.6 ± 2.6 a	47.1 ± 3.7 b	55.8 ± 4.5 ab

Data shown are mean values ± standard errors ($n = 10$ islands per size class). Within each row, differences between numbers followed by the same letter are not statistically significant at $P = 0.05$ (Tukey's test, following one-way ANOVA). Data are from previous studies^{8,22,30,31,45}.

Reciprocal signalling by Notch–Collagen V–CALCR retains muscle stem cells in their niche

Meryem B. Baghdadi^{1,2,3}, David Castel^{4,5}, Léo Machado⁶, So-ichiro Fukada⁷, David E. Birk⁸, Frederic Relaix⁶, Shahragim Tajbakhsh^{1,2*} & Philippos Mourikis^{6*}

The cell microenvironment, which is critical for stem cell maintenance, contains both cellular and non-cellular components, including secreted growth factors and the extracellular matrix^{1–3}. Although Notch and other signalling pathways have previously been reported to regulate quiescence of stem cells^{4–9}, the composition and source of molecules that maintain the stem cell niche remain largely unknown. Here we show that adult muscle satellite (stem) cells in mice produce extracellular matrix collagens to maintain quiescence in a cell-autonomous manner. Using chromatin immunoprecipitation followed by sequencing, we identified NOTCH1/RBPJ-bound regulatory elements adjacent to specific collagen genes, the expression of which is deregulated in Notch-mutant mice. Moreover, we show that Collagen V (COLV) produced by satellite cells is a critical component of the quiescent niche, as depletion of COLV by conditional deletion of the *Col5a1* gene leads to anomalous cell cycle entry and gradual diminution of the stem cell pool. Notably, the interaction of COLV with satellite cells is mediated by the Calcitonin receptor, for which COLV acts as a surrogate local ligand. Systemic administration of a calcitonin derivative is sufficient to rescue the quiescence and self-renewal defects found in COLV-null satellite cells. This study reveals a Notch–COLV–Calcitonin receptor signalling cascade that maintains satellite cells in a quiescent state in a cell-autonomous fashion, and raises the possibility that similar reciprocal mechanisms act in diverse stem cell populations.

Notch activation antagonizes myogenesis by induction of transcriptional repressors (members of the HES/HEY family) and sequestration of the co-activator Mastermind-like 1 from the muscle differentiation factor MEF2C^{10,11}. However, Notch signalling has broader functions in muscle cells, including the maintenance of quiescence^{4,5}. To explore these functions, we carried out chromatin immunoprecipitation followed by sequencing (ChIP-seq) screening¹² and observed that intracellular Notch (NICD) and its downstream effector RBPJ occupied and regulated enhancers proximal to the collagen genes *Col5a1*, *Col5a3*, *Col6a1* and *Col6a2*, which code for collagens that are amongst the most abundant of those produced by satellite cells (Fig. 1a, b and Extended Data Fig. 1a–e). By analysing mouse genetic models with altered Notch activity, we showed that the expression of these collagens tightly correlated with Notch activity in vivo (Extended Data Fig. 2a–e). Moreover, transcriptional induction of *Col5a1* and *Col5a3* by NICD translated to elevated COLV protein levels, specifically the $\alpha 1(\text{V})\alpha 2(\text{V})\alpha 3(\text{V})$ isoform ($\alpha 3$ -COLV), in fetal forelimb (Fig. 1c) and adult hindlimb (tibialis anterior muscle) myogenic cells (Fig. 1d and Extended Data Fig. 2f for $\alpha 3$ -COLV antibody specificity). Furthermore, we isolated collagen-depleted myofibres after treatment with collagenase, to monitor de novo $\alpha 3$ -COLV production. As *Col5a1* and *Col5a3* transcripts are downregulated upon exit from quiescence (Extended Data Figs. 1a, 2g), no $\alpha 3$ -COLV was detected in freshly isolated or activated satellite cells.

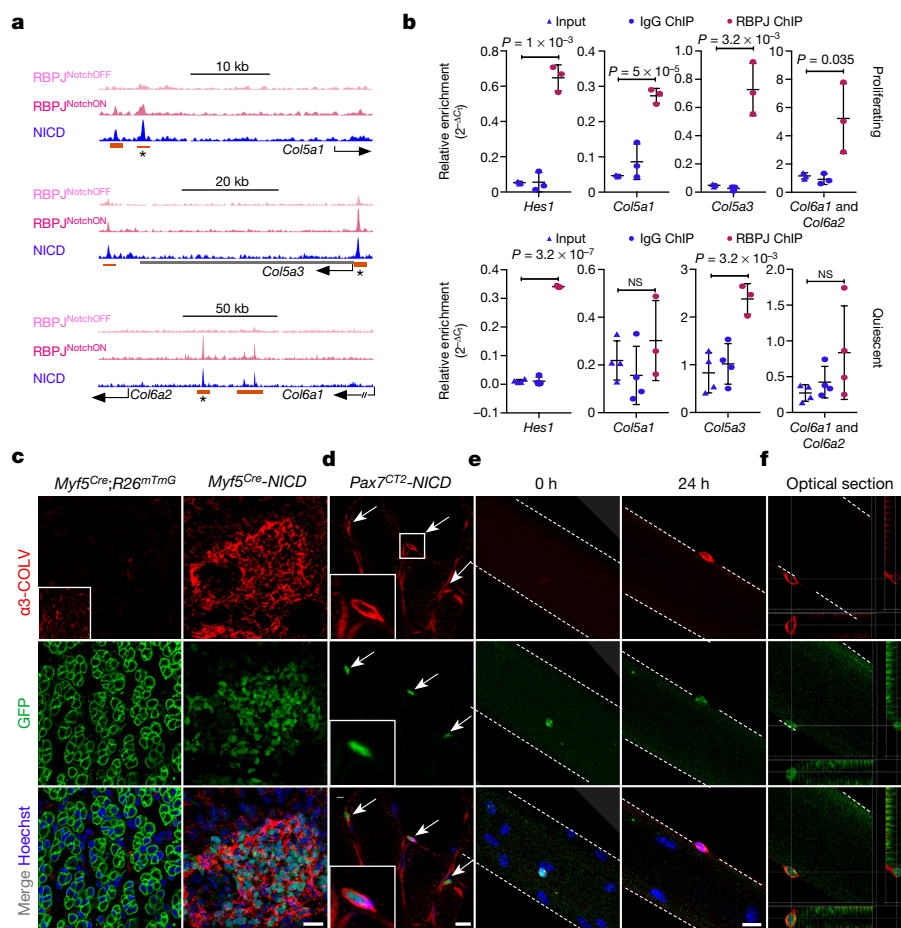
Instead, genetic overexpression of NICD resulted in abundant, newly synthesized $\alpha 3$ -COLV (Fig. 1e, f).

To assess the functional role of COLV, isolated satellite cells were incubated with COLI, COLV or COLVI in the presence of 5-ethynyl-2'-deoxyuridine (EdU) to assess proliferation and stained for PAX7, which marks muscle stem and/or progenitor cells, and the muscle commitment (MYOD) and differentiation (Myogenin) proteins. Only the COLV-complemented medium delayed entry of quiescent cells into the cell cycle (32 h, Fig. 2a), and consequently delayed their amplification and differentiation (72 h, Fig. 2b; 10 days, Extended Data Fig. 3a–c). As previously shown^{4,13}, *Rbpj*^{−/−} cells underwent precocious differentiation and this was partially antagonized by COLV, consistent with the finding that *Col5a1* and *Col5a3* genes are targets of NICD and RBPJ (Fig. 2c, d and Extended Data Fig. 3d–g). Taken together, these results show that COLV, specifically, sustains primary muscle cells in a more stem-like PAX7⁺ state, indicating that COLV could potentially have a role in maintaining the quiescent niche.

To determine whether COLV produced by satellite cells is a functional component of the niche, we generated compound *Tg:Pax7-CreERT2;Col5a1^{flox/flox};R26^{mTmG}* (hereafter referred to as '*Col5a1* cKO') mice, in which COLV was depleted and simultaneously lineage-traced in GFP⁺ satellite cells^{4,14} (Fig. 3a and Extended Data Fig. 4a). Because the $\alpha 1$ -chain of COLV is present in all COLV isoforms (which are trimeric), *Col5a1* deletion produces cells completely lacking COLV protein¹⁴. Unexpectedly—given the general stability of collagens—targeted deletion of *Col5a1* resulted in upregulation of the differentiation marker genes *Myod* (also known as *Myod1*) and *Myog*, and a concomitant reduction of the quiescence marker *Calcr*, as well as *Pax7*, only 18 days after tamoxifen treatment (Fig. 3b). Mutant cells also showed ectopic expression of Myogenin (Fig. 3c), increased 5-bromo-2'-deoxyuridine (BrdU) incorporation (Fig. 3d) and showed a significant decline in PAX7⁺ satellite cells (Fig. 3e). The *Col5a1* cKO cells did not undergo apoptosis (data not shown), but fused to give rise to GFP-marked myofibres (Fig. 3f). Therefore, blocking de novo synthesis of COLV resulted in the spontaneous exit of satellite cells from quiescence, and differentiation, a phenotype reminiscent of Notch loss-of-function^{4,5}.

To investigate the role of *Col5a1* in regeneration, we examined the morphology of tibialis anterior muscles of *Col5a1* cKO mice, 18 days after cardiotoxin-mediated injury (Fig. 3a). Notably, mutant myogenic cells produced smaller nascent myofibres compared to control cells (Fig. 3g, h). Unexpectedly, fewer self-renewing PAX7⁺ cells were observed in the *Col5a1* cKO mice (Fig. 3i) in spite of abundant COLV in regenerating muscle (data not shown), probably produced by the resident fibroblasts, suggesting a cell-autonomous role for *Col5a1*. To investigate self-renewal in a more tractable system, we targeted COLV using short interfering RNA (siRNA) on isolated myofibres in culture in which satellite cells proliferate and self-renew on the myofibre.

¹Department of Developmental & Stem Cell Biology, Institut Pasteur, Paris, France. ²CNRS UMR 3738, Institut Pasteur, Paris, France. ³Sorbonne Universités, UPMC, University of Paris 06, Paris, France. ⁴UMR8203, CNRS, Gustave Roussy, Université Paris-Sud, Université Paris-Saclay, Villejuif, France. ⁵Département de Cancérologie de l'Enfant et de l'Adolescent, Gustave Roussy, Université Paris-Sud, Université Paris-Saclay, Villejuif, France. ⁶INSERM IMRB U955-E10, UPEC, ENVA, EFS, Créteil, France. ⁷Laboratory of Molecular and Cellular Physiology, Graduate School of Pharmaceutical Sciences, Osaka University, Osaka, Japan. ⁸Department of Molecular Pharmacology & Physiology, University of South Florida Morsani College of Medicine, Tampa, FL, USA. *e-mail: shahragim.tajbakhsh@pasteur.fr; philippos.mourikis@inserm.fr



Consistent with our in vivo observations, *Col5a1* knockdown by siRNAs resulted in a marked decrease in the number of the self-renewing PAX7⁺MYOD⁻ cells, compared to scramble control cells (Extended Data Fig. 4b, c). Of note, *Col5a3* siRNA phenocopied *Col5a1* siRNA, which demonstrates that the active triple helix contains α 3-COLV (Extended Data Fig. 4c).

Substrate rigidity and geometry have previously been demonstrated to control stem cell properties, including differentiation and self-renewal^{15,16}. However, we observed that COLV interacted with myogenic cells only when added in the medium, and not when present as a coating substrate (data not shown), which led us to speculate that it acted as a signalling molecule rather than a biomechanical modulator.

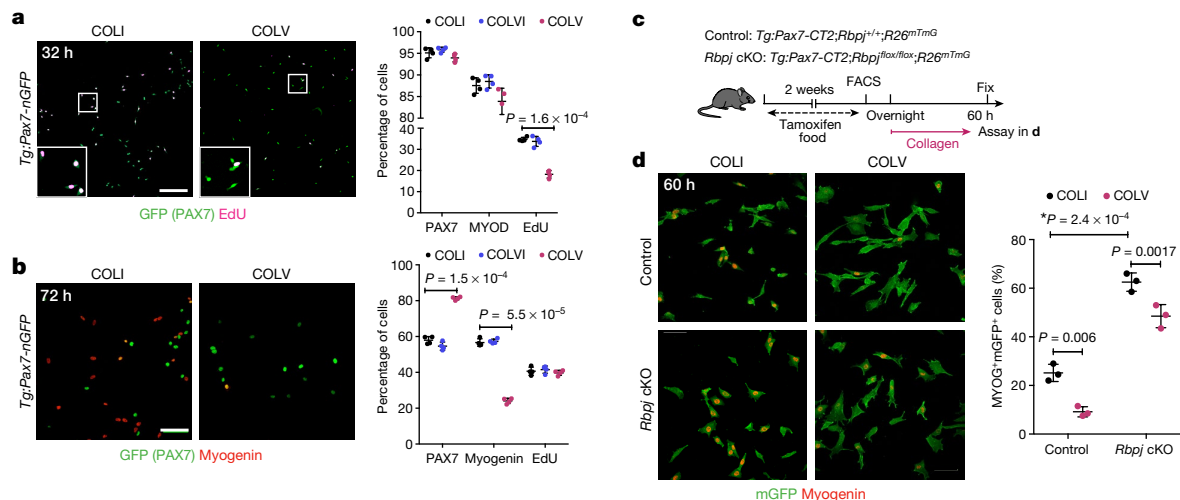


Fig. 2 | COLV delays proliferation and differentiation of satellite cells.

a, EdU pulse (for 2 h) of freshly isolated satellite cells cultured for 32 h: COLI (35%), COLVI (34%) and COLV (18%) (*n* = 4 mice, \geq 250 cells, 2 wells per condition). *Tg:Pax7-nGFP* mice express nuclear (n)GFP driven by *Pax7* regulatory elements. **b**, Immunostaining of freshly isolated satellite cells cultured for 72 h. PAX7: 58%, 55% and 81%; Myogenin: 56%, 57% and 24% for COLI, COLVI and COLV, respectively (*n* = 4 mice, \geq 250

cells, 2 wells per condition). **c**, Experimental scheme for satellite cells plated overnight before collagen treatment. cKO, conditional knockout. **d**, Immunostainings of freshly isolated satellite cells incubated with collagens for 60 h (*n* = 3 mice, \geq 200 cells, 2 wells per condition). Percentage (%) is presented over total GFP⁺ cells. Data are mean \pm s.d.; two-sided paired *t*-test; *, *P* value calculated by two-sided unpaired *t*-test. Scale bars, 50 μ m.

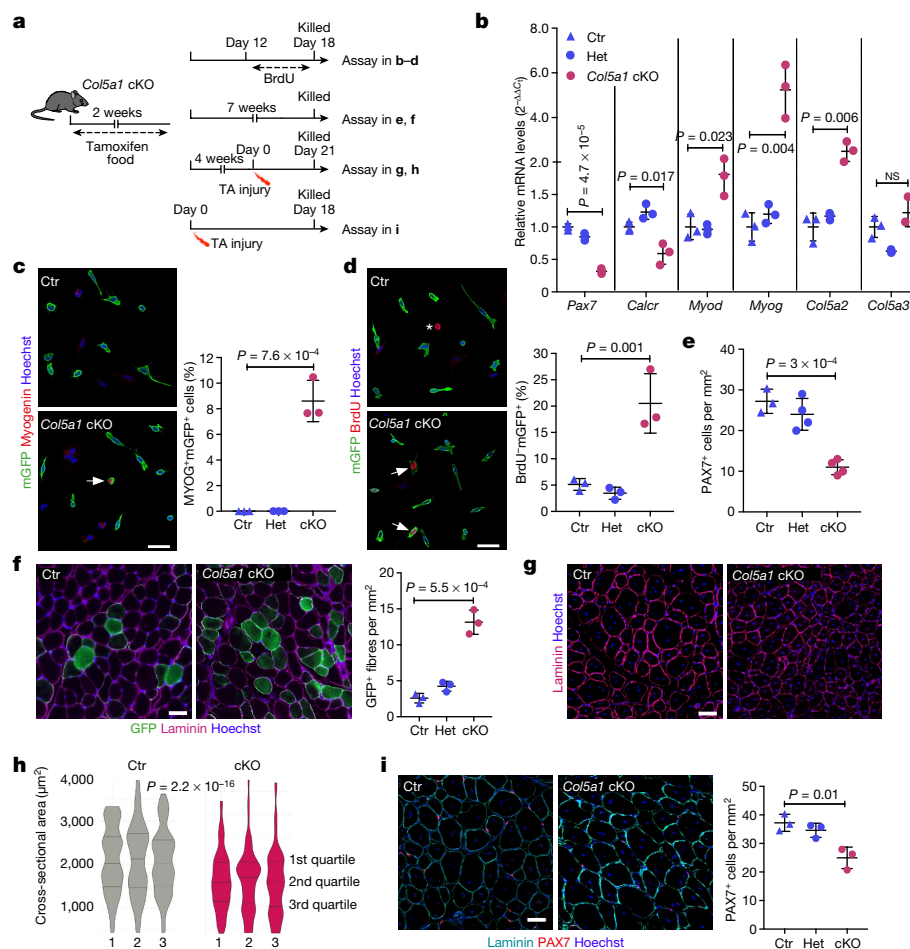


Fig. 3 | Satellite-cell-produced COLV is required in vivo for self-renewal and maintenance of quiescence. **a**, Experimental schemes for control (Tg:Pax7-CT2;Col5a1^{+/+};R26^{mTmG}), heterozygous (Tg:Pax7-CT2;Col5a1^{flox/+};R26^{mTmG}) and conditional knockout (Tg:Pax7-CT2;Col5a1^{flox/flox};R26^{mTmG}) mice. TA, tibialis anterior muscle. **b**, RT-qPCR of satellite cell (Pax7, Calcr) and differentiation (Myod, Myog) markers on Col5a1^{-/-} and control satellite cells isolated by fluorescence-activated cell sorting from resting muscle (n = 3 mice per genotype). Ctr, control; Het, heterozygous; cKO, conditional knockout. **c**, Representative images of membrane-bound GFP⁺ (mGFP) satellite cells from total muscle preparations from control and Col5a1-null mice plated for 12 h. Arrow, mGFP⁺Myogenin⁺ cell (n = 3 mice per genotype, ≥200 cells). **d**, mGFP⁺ satellite cells from total muscle preparations plated for 12 h. Asterisk, non-recombined BrdU⁺ cell; arrows, mGFP⁺BrdU⁺ cells (n = 3 mice per genotype, ≥250 cells). **e**, Satellite cell quantification in quiescent tibialis anterior muscles (seven

weeks after tamoxifen treatment) in control, heterozygous and Col5a1 cKO mice (n = 3 (control) and 4 (heterozygous and cKO) tibialis anterior muscles per genotype). **f**, Immunostaining of sections from control and Col5a1 cKO tibialis anterior muscles seven weeks after tamoxifen treatment (n = 3 mice per genotype). **g**, Immunostaining of sections from control and Col5a1 cKO tibialis anterior muscles 21 days after injury (n = 3 mice per genotype). **h**, Muscle cross-sectional area distribution 21 days after injury (shown as violin plots) was significantly different in control versus Col5a1 cKO tibialis anterior muscles, based on Kruskal–Wallis test (n = 3 mice per genotype, 1,000 fibres analysed per mouse). **i**, Immunostaining of sections 18 days after cardiotoxin injury of control and Col5a1 cKO tibialis anterior muscles (n = 3 mice per genotype). Percentage (%) is presented over total GFP⁺ cells. Data are mean ± s.d.; two-sided unpaired t-test. Scale bars, 50 μm (c, d) and 100 μm (f, g, i).

To identify the cell surface receptor of COLV on satellite cells, we used a myotube-formation assay (see Extended Data Fig. 3b), coupled to inhibitors against known collagen receptors, including Integrins and the RTK receptor DDR1^{17,18}, but these did not obstruct the anti-myogenic activity of COLV (Extended Data Fig. 5a). Because collagens have also previously been shown to bind G-protein coupled receptors (GPCRs)^{19,20}, we focused on Calcitonin receptor (CALCR), which is a GPCR critical for the maintenance of satellite cells²¹. Only cells that expressed CALCR showed decreased proliferation in the presence of COLV (Extended Data Fig. 5b), and Calcr^{-/-} satellite cells isolated from conditional knockout Pax7^{CreERT2};Calcr^{flox/flox} mice failed to respond to COLV treatment (Fig. 4a and Extended Data Fig. 5c–e), demonstrating that CALCR constitutes an essential mediator of the COLV signal (Extended Data Fig. 4e). Accordingly, as CALCR is rapidly cleared after satellite cell activation²¹, COLV had no effect on cultured myogenic cells that had been activated in vivo (three days after injury; Extended Data Fig. 5f). However, we note that addition of COLV on freshly isolated satellite cells appeared to stabilize residual CALCR and

retain Calcr gene expression, thus allowing their prolonged interaction (Extended Data Fig. 5g–i). In summary, we show that CALCR is a critical mediator of the effect of COLV on maintaining quiescence and on the stemness properties of satellite cells.

To date, it has been assumed that CALCR in satellite cells is activated by circulating calcitonin peptide hormones, which are principally expressed by parafollicular thyroid cells; this points to systemic regulation of stem cell quiescence. Based on our findings, we reasoned that COLV serves as a local ligand for the CALCR receptor. Indeed, on-cell enzyme-linked immunosorbent assay experiments showed that COLV—but not COLI—selectively bound to cells expressing CALCR (Fig. 4b). Notably, this binding was functional as COLV—but not COLI—displayed rapid activation kinetics and upregulation of levels of intracellular cAMP, which is a downstream reporter of CALCR activation²² (Fig. 4c, d and Extended Data Fig. 6a). In vitro binding assays using the extracellular domain of CALCR did not result in robust interaction with COLV (data not shown). Therefore, we propose that binding of COLV to CALCR requires a specific configuration of the

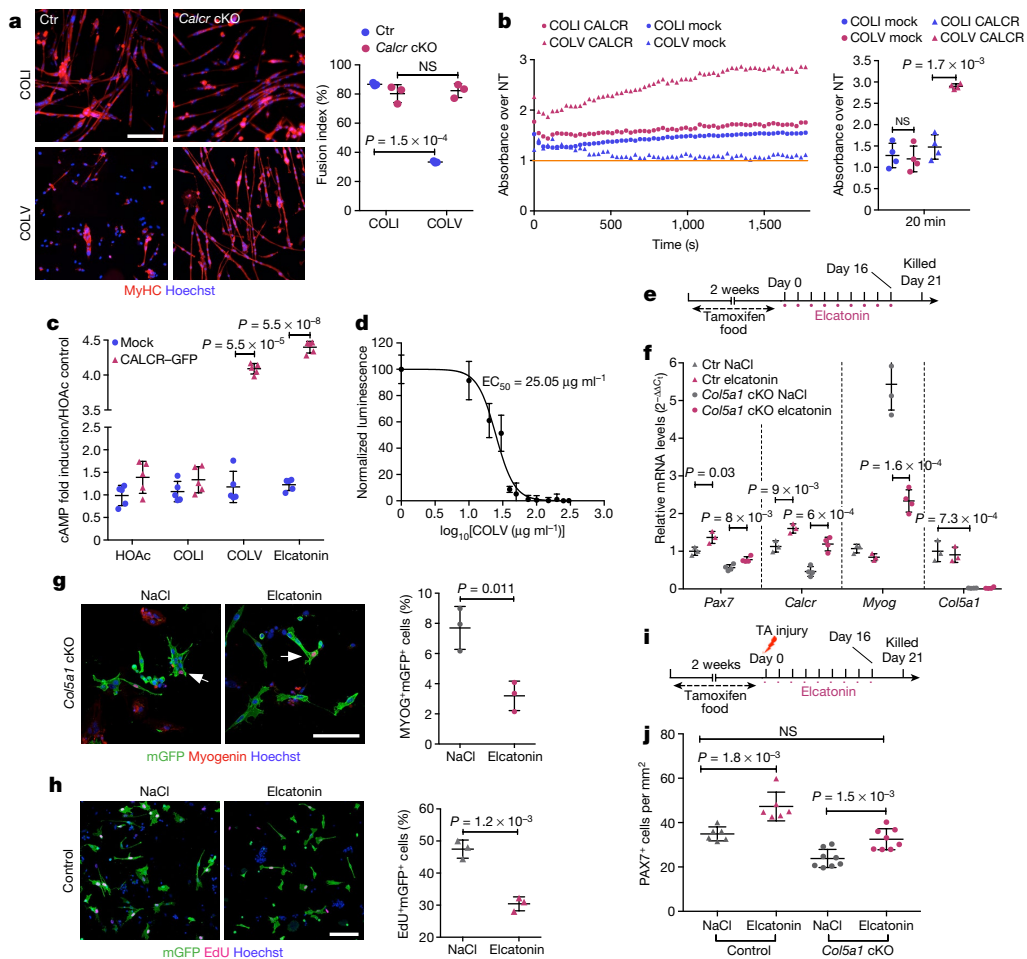


Fig. 4 | Interaction of COLV with satellite cells is mediated by CALCR.

a, Control (Ctr; *Pax7*^{CT2/+}; *Calcr*^{+/+}; *R26*^{stop-YFP}) and *Calcr*-deficient (*Calcr* cKO; *Pax7*^{CT2/+}; *Calcr*^{flx/flx}; *R26*^{stop-YFP}) satellite cells incubated for 10 days with COLI or COLV and immunostained for differentiation ($n = 3$ mice, ≥ 250 cells). **b**, Binding assay of COLV and CALCR by colorimetric on-cell enzyme-linked immunosorbent assay based on the measurements of horseradish peroxidase absorbance. Runs test P value < 0.0001 . Results presented as ratio of absorbance over non-treated cells (NT, orange line = 1) at 20 min of horseradish peroxidase development. **c**, cAMP measurements of *Calcr*-transduced C2C12 cells after three hours of treatment with acetic acid (HOAc), COLI, COLV or elcatonin. Graph represents fold cAMP induction over average of mock cells treated with HOAc ($n = 4$ assays). **d**, Dose-response curve of fold cAMP concentration in *Calcr*-transduced C2C12 cells treated for 3 h with increasing concentrations of COLV. Half-maximal effective concentration (EC₅₀) = 25.05 μg ml⁻¹ ($n = 4$ independent assays). **e**, Experimental

scheme of tamoxifen and elcatonin administration to *Col5a1* cKO mice and their corresponding control mice. **f**, RT-qPCR of satellite cells (*Pax7*, *Calcr*) and differentiation (*Myog*) markers on *Col5a1* cKO mutant mice and control mice ($n = 3$ mice per condition) treated with elcatonin or saline. **g**, Representative images of mGFP⁺ satellite cells from total muscle preparations from *Col5a1*-null mice injected with saline or elcatonin, plated for 12 h. Arrows, mGFP⁺MYOG⁺ cells ($n = 3$ mice per condition, ≥ 200 cells). **h**, EdU (for 2 h) and mGFP staining of satellite cells from total muscle preparations from control mice treated with saline or elcatonin, plated for 36 h. Asterisk, mGFP⁺EdU⁺ cell ($n = 3$ mice per genotype, ≥ 400 cells). **i**, Experimental scheme of tamoxifen and elcatonin administration to control and *Col5a1* cKO mice. **j**, PAX7⁺ cells on tibialis anterior sections 21 days after injury in mice treated with saline or elcatonin ($n = 6$ mice for control and 8 mice for cKO, per treatment). Percentage (%) is presented over total GFP⁺ cells. Data are mean \pm s.d.; **a–c**, two-sided paired t -test; **f–j**, two-sided unpaired t -test. Scale bars, 25 μm.

receptor, possibly involving the extracellular loops or co-factors. Taken together, these data demonstrate that COLV physically and functionally interacts with CALCR.

In this study, we showed that blocking COLV production from satellite cells resulted in rupture of quiescence and impaired self-renewal in vivo. Combined with our ex vivo results, the similarity of these phenotypes to Notch and CALCR signalling abrogation points to a cell-autonomous Notch–COLV–CALCR axis that sustains muscle stem cells in their niche. Consistent with this notion, administration of the CALCR ligand elcatonin to control and *Col5a1*-null mice resulted in upregulation of the stem cell markers *Pax7* and *Calcr*, indicating that the injected ligand was readily delivered to the quiescent satellite cells (Fig. 4e, f). Notably, elcatonin mitigated the precocious *Myog* transcription and protein expression levels in *Col5a1* mutant cells (Fig. 4f, g). Elcatonin also prolonged the G0-to-S transition of control satellite cells exiting quiescence (Fig. 4h), which suggests that hyperactivation of

CALCR could drive cells into a deeper, more dormant-like quiescent state marked by higher *Pax7* expression²³. Therefore, CALCR activity appears to control quiescence quantitatively, shown by the loss of satellite cells in the absence of ligand COLV, and qualitatively, shown by the presence of dormant-like satellite cells upon hyperactivation. Elcatonin restored the number of PAX7⁺ satellite cells in regenerating *Col5a1* cKO muscles to wild-type levels (Fig. 4i, j), and in an ex vivo self-renewal reserve-cell model (Extended Data Fig. 6b, c). Therefore, we show that endogenous calcitonin levels are not sufficient to maintain *Col5a1*-null satellite cells, and that exogenous administration of a calcitonin derivative rescued the defects, probably via the activation of CALCR.

Here we describe a self-sustained signalling cascade orchestrated by the Notch pathway and propagated by the extracellular matrix of the immediate skeletal muscle stem cell niche (Extended Data Fig. 7). We propose that Notch acts as a sensor of the homeostatic environment

by reinforcing the niche with active COLV that provides cell-autonomous signals and maintains stem cell quiescence. Upon disruption of the niche and physical separation of the ligands, Notch signalling is sharply downregulated and stem cells exit quiescence^{4,24}. This halts further production of COLV and thus favours satellite cell activation, as shown in our model (Extended Data Fig. 7). It would be of interest to investigate whether the Notch–COLV–CALCR signalling cascade described here applies to stem cells in other tissues and organisms, in which an extracellular matrix protein produced by the stem cell can act as a local ligand for cell-autonomous stability of the niche through a GPCR. The regulatory mechanism that we identify provides a framework to construct a more complete view of the stem cell niche, and to manipulate stem cell behaviour in a therapeutic context.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0144-9>.

Received: 16 January 2017; Accepted: 6 April 2018;

Published online 23 May 2018.

- Raymond, K., Deugnier, M. A., Faraldo, M. M. & Glukhova, M. A. Adhesion within the stem cell niches. *Curr. Opin. Cell Biol.* **21**, 623–629 (2009).
- Moore, K. A. & Lemischka, I. R. Stem cells and their niches. *Science* **311**, 1880–1885 (2006).
- Watt, F. M. & Huck, W. T. Role of the extracellular matrix in regulating stem cell fate. *Nat. Rev. Mol. Cell Biol.* **14**, 467–473 (2013).
- Mourikis, P. et al. A critical requirement for notch signaling in maintenance of the quiescent skeletal muscle stem cell state. *Stem Cells* **30**, 243–252 (2012).
- Bjornson, C. R. et al. Notch signaling is necessary to maintain quiescence in adult muscle stem cells. *Stem Cells* **30**, 232–242 (2012).
- Rozo, M., Li, L. & Fan, C. M. Targeting β 1-integrin signaling enhances regeneration in aged and dystrophic muscle in mice. *Nat. Med.* **22**, 889–896 (2016).
- Cheung, T. H. et al. Maintenance of muscle stem-cell quiescence by microRNA-489. *Nature* **482**, 524–528 (2012).
- Zismanov, V. et al. Phosphorylation of eIF2 α is a translational control mechanism regulating muscle stem cell quiescence and self-renewal. *Cell Stem Cell* **18**, 79–90 (2016).
- Chakkalakal, J. V., Jones, K. M., Basson, M. A. & Brack, A. S. The aged niche disrupts muscle stem cell quiescence. *Nature* **490**, 355–360 (2012).
- Shen, H. et al. The Notch coactivator, MAML1, functions as a novel coactivator for MEF2C-mediated transcription and is required for normal myogenesis. *Genes Dev.* **20**, 675–688 (2006).
- Busas, M. F., Kabak, S. & Kadesch, T. The Notch effector Hey1 associates with myogenic target genes to repress myogenesis. *J. Biol. Chem.* **285**, 1249–1258 (2010).
- Castel, D. et al. Dynamic binding of RBPJ is determined by Notch signaling status. *Genes Dev.* **27**, 1059–1071 (2013).
- Vasyutina, E. et al. RBP-J (Rbbsuh) is essential to maintain muscle progenitor cells and to generate satellite cells. *Proc. Natl Acad. Sci. USA* **104**, 4443–4448 (2007).
- Sun, M. et al. Targeted deletion of collagen V in tendons and ligaments results in a classic Ehlers–Danlos syndrome joint phenotype. *Am. J. Pathol.* **185**, 1436–1447 (2015).
- Gilbert, P. M. et al. Substrate elasticity regulates skeletal muscle stem cell self-renewal in culture. *Science* **329**, 1078–1081 (2010).
- Yennek, S., Burute, M., Théry, M. & Tajbakhsh, S. Cell adhesion geometry regulates non-random DNA segregation and asymmetric cell fates in mouse skeletal muscle stem cells. *Cell Reports* **7**, 961–970 (2014).
- Leitinger, B. Transmembrane collagen receptors. *Annu. Rev. Cell Dev. Biol.* **27**, 265–290 (2011).
- Vogel, W., Gish, G. D., Alves, F. & Pawson, T. The discoidin domain receptor tyrosine kinases are activated by collagen. *Mol. Cell* **1**, 13–23 (1997).
- Paavola, K. J., Sidik, H., Zuchero, J. B., Eckart, M. & Talbot, W. S. Type IV collagen is an activating ligand for the adhesion G protein-coupled receptor GPR126. *Sci. Signal.* **7**, ra76 (2014).
- Luo, R. et al. G protein-coupled receptor 56 and collagen III, a receptor–ligand pair, regulates cortical development and lamination. *Proc. Natl Acad. Sci. USA* **108**, 12925–12930 (2011).
- Yamaguchi, M. et al. Calcitonin receptor signaling inhibits muscle stem cells from escaping the quiescent state and the niche. *Cell Reports* **13**, 302–314 (2015).
- Evans, B. N., Rosenblatt, M. I., Mnayer, L. O., Oliver, K. R. & Dickerson, I. M. CGRP-RCP, a novel protein required for signal transduction at calcitonin gene-related peptide and adrenomedullin receptors. *J. Biol. Chem.* **275**, 31438–31443 (2000).
- Rocheteau, P., Gayraud-Morel, B., Siegl-Cachedenier, I., Blasco, M. A. & Tajbakhsh, S. A subpopulation of adult skeletal muscle stem cells retains all template DNA strands after cell division. *Cell* **148**, 112–125 (2012).
- Mourikis, P. & Tajbakhsh, S. Distinct contextual roles for Notch signalling in skeletal muscle stem cells. *BMC Dev. Biol.* **14**, 2 (2014).
- Machado, L. et al. *In situ* fixation redefines quiescence and early activation of skeletal muscle stem cells. *Cell Reports* **21**, 1982–1993 (2017).
- Mourikis, P., Gopalakrishnan, S., Sambasivan, R. & Tajbakhsh, S. Cell-autonomous Notch activity maintains the temporal specification potential of skeletal muscle stem cells. *Development* **139**, 4536–4548 (2012).

Acknowledgements We thank H. Stunnenberg for the ChIP-seq and RNA sequencing data; D. Castro for the RBPJ ChIP protocol; D. Greenspan for the anti- α 3-COLV antibody and Col5a3-knockout muscle samples; C. Moali for the SPR assay; F. Auradé and the Protein Core Facility, Institut Curie, for the production of CalcR proteins; K. Ding for the 7h DDR1 inhibitor; F. Ruggiero for suggesting the on-cell enzyme-linking immunosorbent assay experiment; and the Cytometry platforms of Institut Pasteur and IMRB, Inserm U955, Creteil. F.R. was funded by the Association Française contre les Myopathies via TRANSLAMUSCLE (PROJECT 19507), Agence Nationale pour la Recherche grant Satnet (ANR-15-CE13-0011-01) and RHU CARMMA (ANR-15-RHUS-0003). S.T. was funded by Institut Pasteur, Centre National pour la Recherche Scientifique and the Agence Nationale de la Recherche (Laboratoire d'Excellence Revive, Investissement d'Avenir; ANR-10-LABX-73) and the European Research Council (Advanced Research Grant 332893). M.B.B. was funded by the Doctoral School grant and Fondation pour la Recherche Médicale.

Reviewer information Nature thanks I. Conboy, G. Kardon and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.B.B., S.T. and P.M. proposed the concept, designed experiments and wrote the manuscript, F.R. oversaw revisions, and S.T. funded most of the study. P.M. and D.C. conducted initial experiments on enhancer analysis. D.C. and L.M. performed and analysed ChIP experiments. M.B.B. performed the remaining experiments and, together with P.M., analysed the data. S.F. and D.E.B. provided mouse models.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0144-9>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0144-9>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to S.T. or P.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Mouse strains. Mouse lines used in this study have been described and provided by the corresponding laboratories: *Myf5^{Cre}* mice²⁷, *Pax7^{CreERT2}* mice²⁸ (used to recombine *R26^{stop-NICD}* allele), *R26^{stop-NICD-nGFP}* mice²⁹, *R26^{mTmG}* mice³⁰ (ROSA 26 gene trap with membrane-Tomato floxed/membrane-GFP), *Rbpj^{fllox/flox}* mice³¹, *Pax7^{CT2/+}*; *Calcr^{fllox/flox}*; *R26^{stop-YFP/stop-YFP}* mice²¹ (triple mutant mice provided by S.F.) and *Col5a1^{fllox/flox}* mice³². *Tg:Pax7-CreERT2* (used to recombine *Rbpj* and *Col5a1*) and *Tg:Pax7-nGFP* lines have previously been described^{4,33}. All adult mice analysed were between 8 and 12 weeks old. Animals were handled according to national and European community guidelines, and protocols were approved by the ethics committee at Institut Pasteur and the French Ministry.

Muscle injury, tamoxifen, BrdU and elcatonin administration. For muscle injury, *Tg:Pax7-CreERT2;Col5a1^{fllox/flox};R26^{mTmG}* mice and their corresponding controls were anaesthetized with 0.5% Imalgene/2% Rompun and the tibialis anterior muscle was injected with 50 μ l of cardiotoxin (10 μ M; Latoxan). *Tg:Pax7-CreERT2;Rbpj^{fllox/flox};R26^{mTmG}* mice and their corresponding controls were injected intraperitoneally with tamoxifen three times (250 to 300 μ l, 20mg/ml; Sigma T5648; diluted in sunflower seed oil/5% ethanol). *Pax7^{CreERT2};Calcr^{fllox/flox};R26^{stop-YFP}* mice and their corresponding controls were injected intraperitoneally with tamoxifen twice (1 mg per 5 g of body weight) and euthanized two weeks later. *Pax7^{CreERT2};R26^{stop-NICD-ires-nGFP}* and *Tg:Pax7-CreERT2;Col5a1^{fllox/flox};R26^{mTmG}* mice and their corresponding controls were fed a diet containing tamoxifen for one and two weeks, respectively (Envigo, TD55125). Six days before being euthanized, *Tg:Pax7-CreERT2;Col5a1^{fllox/flox};R26^{mTmG}* mice and their corresponding controls were given the thymidine analogue BrdU (0.5 mg/ml, #B5002; Sigma) in the drinking water supplemented with sucrose (25 mg/ml). Elcatonin (2.5 ng per g of mouse, final concentration in 0.9% NaCl; Mybiosource, MBS143228) was injected subcutaneously eight times, every other day. Comparisons were done between age-matched littermates using 8–12-week-old mice.

Muscle enzymatic dissociation and stem cell isolation. Adult and fetal limb muscles were dissected, minced and incubated with a mix of Dispase II (Roche, 04942078001) 3 U/ml, collagenase A (Roche, 11088793001) 100 μ g/ml and DNase I (Roche, 11284932001) 10 mg/ml in Hank's Balanced Salt Solution (Gibco) supplemented with 1% penicillin–streptomycin (PS; Gibco) at 37 °C at 60 r.p.m. in a shaking water bath for 2 h. The muscle suspension was successively filtered through 100- μ m and 70- μ m cell strainers (Miltenyi, 130-098-463 and 130-098-462) and then spun at 50g for 10 min at 4 °C to remove large tissue fragments. The supernatant was collected and washed twice by centrifugation at 600g for 15 min at 4 °C. Before fluorescence-activated cell sorting (FACS), the final pellet was resuspended in cold Dulbecco's modified Eagle's medium (DMEM) and 1% PS supplemented with 2% fetal bovine serum (FBS), and the cell suspension was filtered through a 40- μ m strainer. Satellite cells were sorted with Aria III (BD Biosciences) using either the GFP (*Tg:Pax7-nGFP* or *Tg:Pax7-CreERT2;Rbpj^{fllox/flox};R26^{mTmG}* or *Tg:Pax7-CreERT2;Col5a1^{fllox/flox};R26^{mTmG}*) or the YFP (*Pax7^{CT2};Calcr^{fllox/flox};R26^{stop-YFP}*) cell markers. Isolated, mononuclear cells were collected in DMEM/1% PS/2% FBS. Enzymatically dissociated muscle was also plated directly without FACS on Matrigel-coated dishes (Corning, 354248; 30 min at 37 °C), and fixed 12 h later with 4% paraformaldehyde (PFA)/PBS. Cells were immunostained following the protocol described above.

Chromatin immunoprecipitation. *Cultured myoblasts.* Satellite cells were isolated from adult *Tg:Pax7-nGFP* mice and plated on dishes, coated with Delta-like 1, for 72 h to maintain active Notch signalling, as previously described^{14,34}. Cells were then processed for ChIP using a dual cross-linking protocol³⁵, with slight modifications. In brief, cells were fixed on the dish with 2 mM di(N-succinimidyl) glutarate (Sigma, 80424) in PBS for 45 min at room temperature. After two washes with PBS, cells were re-fixed with 1% formaldehyde/PBS for 10 min at room temperature, before quenching the reaction with 1/20 volume of 2.5 M glycine for 5 min at room temperature. The cells were then collected with a cell scraper in PBS supplemented with 1% BSA and protease inhibitors (Roche, 11697498001), and collected by spinning. Cell lysis and chromatin isolation were done using the Ideal ChIP-seq kit for histones (Diagenode, C01010051). Chromatin was sheared using a Bioruptor Pico (Diagenode B01060001) with 10 cycles of 30 s on/off sonication. The samples were prepared in triplicates from different plates. Primary myogenic cells (2×10^6) were used per ChIP and 2×10^4 cells were used per input. The immunoprecipitations were performed following the manufacturer's guidelines using 6 μ l of anti-RBPJ antibody (Cell Signalling, #5313) or 1.5 μ l of rabbit control IgG antibody (Diagenode, C15410206) in a final volume of 300 μ l per ChIP. The purification of the immunoprecipitated DNA was performed using DiaPure columns (Diagenode, C03040001). RT-qPCR was performed using FastStart Universal SYBR Green Master mix (Roche, 04913914001) and analysis was performed using the $2^{-\Delta\Delta C_t}$ method³⁶ normalized to the Neg16 region.

Quiescent satellite cells. Satellite cells were isolated from adult *Tg:Pax7-nGFP* mice using in situ fixation to preserve Notch signalling from dissociation-induced downregulation²⁵. Cells were fixed as above in 2 mM di(N-succinimidyl)

glutarate for 45 min, followed by 10 min with 1% formaldehyde at room temperature. Cell lysis and chromatin isolation were performed using Auto-TrueMicrochip kit (Diagenode, C01010140). Chromatin was sheared as above with 10 cycles of 30 s on/off sonication using a Bioruptor Pico. Two hundred thousand cells were used per ChIP and 2×10^3 per input and IPs were performed using 2 μ l of anti-RBPJ antibody (Cell Signalling, 5313) or 0.5 μ l of rabbit control IgG antibody following the manufacturer's guidelines. Immunoprecipitated chromatin preparations and input were purified using the Auto IPure kit v2 (Diagenode). RT-qPCR was performed using FastStart Universal SYBR Green Master mix (Roche, 04913914001) and analysis was performed using the $2^{-\Delta\Delta C_t}$ method³⁶ normalized to the Neg16 region. Primers used for ChIP-qPCR are listed in Supplementary Table 1.

Cell culture and collagen incubation. Satellite cells isolated by FACS were plated at 3×10^3 cells per cm² on ibi-treated μ -slides (Ibidi, 80826) pre-coated with 0.1% gelatin for 2 h at 37 °C. Cells were cultured in satellite cell growth medium containing DMEM (Gibco) supplemented with F12 (50:50; Gibco), 1% PS, 20% FBS (Gibco) and 2% Ultrosor (Pall; 15950-017) at 37 °C, 3% O₂, 5% CO₂ for the indicated time. Twelve hours after plating, collagens (COLI rat tail, BD Biosciences, 354236; COLV human placenta, Sigma, C3657; COLVI human placenta, AbD Serotec 2150-0230) resuspended in HOAc acid at 1 mg/ml, were added to the culture medium at a final concentration of 50 μ g/ml and cells were fixed with 4% PFA for 10 min at room temperature. To assess proliferation, cells were pulsed with the thymidine analogue EdU, 1×10^{-6} M at 2 h before fixation (ThermoFisher Click-iT Plus EdU kit, C10640). Inhibitors used: Obtustatin (Integrin α 1 β 1, Tocris, 4664, 100 nM), TC-I 15 (Integrin α 2 β 1 Tocris, 4527, 100 μ M), RGDS peptide (all Integrins, Tocris, 3498, 100 μ M), 7rh³⁷ (DDR1, a gift from K. Ding, 20 nM).

Muscle fixation and histological analysis. Embryo forelimbs were fixed in 4% PFA/0.1% Triton for 2 h, washed overnight with $1 \times$ PBS, immersed in 20% sucrose/PBS overnight, embedded in OCT, frozen in liquid nitrogen and sectioned transversely at 12–14 μ m. Isolated tibialis anterior muscles were immediately frozen in liquid-nitrogen-cooled isopentane and sectioned transversely at 8 μ m. For PAX7 staining on adult tibialis anterior muscle, sections were post-fixed with 4% PFA, 15 min at room temperature. After 3 washes with $1 \times$ PBS, antigen retrieval was performed by incubating sections in boiling 10 mM citrate buffer pH 6 for 10 min. Sections were then blocked, permeabilized and incubated with primary and secondary antibodies as described in 'Immunostaining on cells, sections and myofibres'.

Single myofibre isolation and siRNA transfection. Single myofibres were isolated from extensor digitorum longus muscles following the previously described protocol³⁸. In brief, extensor digitorum longus muscles were dissected and incubated in 0.1% w/v collagenase (Sigma, C0130)/DMEM for 1 h in a 37 °C shaking water bath at 40 r.p.m. Following enzymatic digestion, mechanical dissociation was performed to release individual myofibres that were then transferred to serum-coated Petri dishes. Single myofibres were transfected with *Col5a1* siRNA, *Col5a3* siRNA (Dharmacon SMARTpool *Col5a1* (12831) L-044167-01 and *Col5a3* (53867) L-048934-01-0005) or scramble siRNA (Dharmacon ON-TARGETplus Non-targeting siRNA #2 D-001810-02-05) at a final concentration of 200 nM, using Lipofectamine 2000 (ThermoFisher, 11668) in Opti-MEM (Gibco). Four hours after transfection, 6 volumes of fresh satellite cell growth medium were added and fibres were cultured for 72 h at 37 °C, 3% O₂. Myofibres were fixed for 15 min in 4% PFA before immunostaining for proliferation, differentiation and self-renewal markers³⁹.

Immunostaining on cells, sections and myofibres. Following fixation, cells and myofibres were washed three times with PBS, then permeabilized and blocked at the same time in buffer containing 0.25% Triton X-100 (Sigma), 10% goat serum (Gibco) for 30 min at room temperature. For BrdU immunostaining, cells were unmasked with DNaseI (1,000 U/ml, Roche, 04536282001) for 30 min at 37 °C. Cells and fibres were then incubated with primary antibodies (Supplementary Table 2) for 4 h at room temperature. Samples were washed with $1 \times$ PBS three times and incubated with Alexa-conjugated secondary antibodies (Life Technologies, 1/1,000) and Hoechst 33342 (Life Technologies, 1/5,000) for 45 min at room temperature. EdU staining was chemically revealed using the Click-iT Plus kit according to manufacturer's recommendations (Life Technologies, C10640). For collagen staining, the myofibres and the muscle sections were incubated with 0.1% Triton X-100 for 30 min at room temperature. Myofibres and sections were then washed 3×10 min and incubated with 10% goat serum in PBS for 30 min. After one wash, samples were incubated with primary antibodies and secondary antibodies as described in Supplementary Table 2. Confocal images were acquired with a Leica SPE microscope and Leica Application Suite or with Zeiss LSM 700 microscope and Zen Blue 2.0 software. Three-dimensional images were reconstructed from confocal Z-stacks using Imaris software. The Section view function was used to inspect the environment of the satellite cells by showing the cut in the x, y and z axes.

Reserve cell cultures. Enzymatically dissociated muscles were plated in gelatin-coated dishes (1/30 of total mouse muscles per cm²) in the satellite cell growth

medium described above. When myotube formation was detected (day 7 to 10), recombination was induced by addition of 4-hydroxytamoxifen (4-OHT; Sigma, H6278) at final concentration of 1 μ M every other day. Seven days later, 4-OHT-containing medium was replaced every other day with fresh medium containing elcatonin (0.1 U/ml), for an additional 10 days. To assess proliferation, cells were pulsed with 1×10^{-6} M EdU for 6 h before fixation (10 min, 4% PFA). Reserve cells were defined by immunofluorescence as PAX7⁺EdU⁺ cells³⁹. For each medium change, only half of the conditioned medium was removed and replaced by an equal volume of fresh medium.

Construction of luciferase reporters and luciferase assays. For the generation of luciferase reporters, candidate enhancers of *Col5a1*, *Col5a3*, *Col6a1* and *Col6a2* (a shared enhancer), and *Hey1* were amplified by PCR from genomic DNA of C2C12 cells. The enhancers were then cloned into the firefly-luciferase pGL3-Basic vector (Promega, E1751) upstream of a minimal thymidine kinase promoter. The sequences of enhancers are listed in Supplementary Table 3. Transfected cells (Lipofectamine LTX, Life Technologies, 15338030) were lysed and luciferase signal was scored using the Dual-Luciferase Reporter Assay System (Promega, E1910). For normalization, *Renilla* luciferase (pCMV-Renilla) was transfected at 1:20 ratio relative to firefly-luciferase constructs.

RNA isolation and RT-qPCR. Total RNA was extracted from satellite cells isolated by FACS using QIAGEN mini RNeasy kit and reverse transcribed using SuperScript III (Invitrogen, 18080093), according to manufacturer's instructions. RT-qPCR was performed using FastStart Universal SYBR Green Master mix (Roche, 04913914001) and analysis was performed employing the $2^{-\Delta\Delta C_t}$ method and using the average of the control values as a reference³⁶. Specific forward and reverse primers used in this study are listed in Supplementary Table 1.

Stable cell line manipulations. The mouse myoblast cell line C2C12 was cultured in DMEM/ 20% FBS/ 1% PS at 37°C, 5% CO₂.

Notch activation. Notch activation was achieved by plating cells on DLL1-coated dishes or by doxycycline-inducible Notch constructs, as previously described¹².

Calcr retrovirus preparation and transduction. Calcitonin receptor C1a-type (pMXs-Calcr-C1a-IRES-GFP) and mock control (pMXs-IRES-GFP) retrovirus vectors were prepared as previously described^{21,40}. In brief, 48 h after transfection of Platinum-E cells the supernatant was recovered and used to transduce C2C12 cells. Two days later stably labelled GFP⁺ C2C12 cells were isolated by FACS. All stable cell lines used in this study are negative for mycoplasma contamination.

Quantification of cAMP. Transduced mock (IRES-GFP) and *Calcr* (Calcr-C1a-IRES-GFP) C2C12 cells were isolated by FACS based on GFP expression and seeded on 0.1% gelatin-coated, white culture 96-well plates (Falcon, 353296) at 3×10^3 cells per well. After overnight culture, the cells were incubated with the complete induction medium containing DMEM/1% PS/500 μ M isobutyl-1-methylxanthine (Sigma, 17018)/100 μ M 4-(3-butoxy-4-methoxy-benzyl)imidazolidone (Ro 20-1724 Sigma, B8279)/MgCl₂ 40 mM, collagen, solvent HOAc or elcatonin (0.1 U/ml) for 3 h. The amount of intracellular cAMP was measured using cAMP-Glo Max Assay (Promega, V1681) following the manufacturer's protocol. Luminescence was quantified with FLUOstar OPTIMA (BMG Labtech). The EC₅₀ value was determined with GraphPad Prism software using a sigmoid dose-response curve (variable slope).

Biotinylation of collagens. Commercial collagen proteins (COLI rat tail, BD Biosciences, 354236; COLV human placenta, Sigma, C3657) were biotinylated using the Pierce EZ-Link Biotinylation Kit, with slight modifications. In brief, 20 μ l of 1 M HEPES was added to 0.5 ml of 1 mg/ml collagen dissolved in 0.5 M HOAc. Then, 20 μ l of 100 mM biotin reagent were added and incubated at room temperature for 1.5 h. Biotinylated collagens were next dialysed in 25 mM HEPES, 2.5 M CaCl₂, 125 mM NaCl, 0.005% Tween (Slide-A-Lyze MINI Dialysis Device, Thermo Fisher 88401) overnight at 4°C.

On-cell enzyme-linked immunosorbent assay. Transduced mock and *Calcr* C2C12 cells were seeded on a clear-bottom 96-well plate (TPP, 92096) at a density

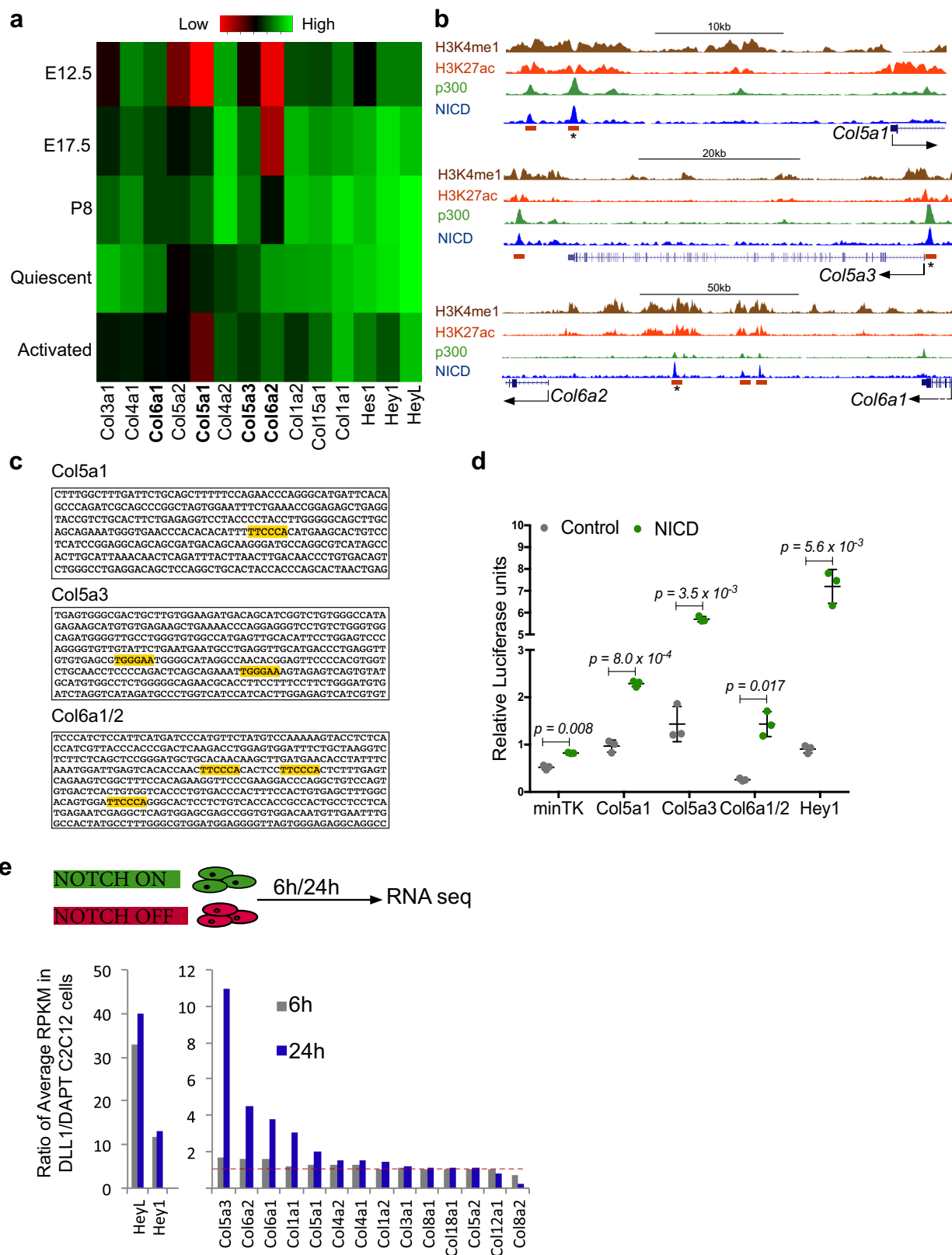
of 3×10^3 cells per well. After overnight culture, cells were treated with 50 μ g/ml of biotinylated collagens for 2 h and fixed with 4% PFA/PBS for 15 min. After $3 \times$ PBS washes, cells were blocked with a solution containing 10% goat serum, 2% BSA, PBS for 1 h at room temperature, washed and incubated for 1 h at room temperature with goat anti-mouse biotin-HRP antibody (Jackson, 1/1000, 115-035-003). After $3 \times$ PBS washes, the HRP signal was developed by addition of 3,3',5,5'-tetramethylbenzidine (1-Step Ultra TMB-ELISA, Sigma, 34028). HRP substrate and absorbance at 650 nm was measured once every 30 s for 30 min with FLUOstar OPTIMA (BMG Labtech). The signal was normalized to the background signal (no secondary antibody) and to the number of cells assessed by Janus green staining (Abcam, ab111622).

Statistical analysis. No statistical methods were used to predetermine sample size. The investigators were not blinded to allocation during experiments and outcome assessment. No animal has been excluded from analysis and no randomization method has been applied in this study. For comparison between two groups, two-tailed paired and unpaired Student's *t*-tests were performed to calculate *P* values and to determine statistically significant differences (see legends of Figs. 1–4). Additional specific statistical tests are detailed in legends of Figs. 1–4. All experiments have been done twice with the same results. All statistical analyses were performed with Excel software or GraphPad Prism software; Kruskal–Wallis test was performed in R.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. All data that support the findings of this study are available from the corresponding authors upon request.

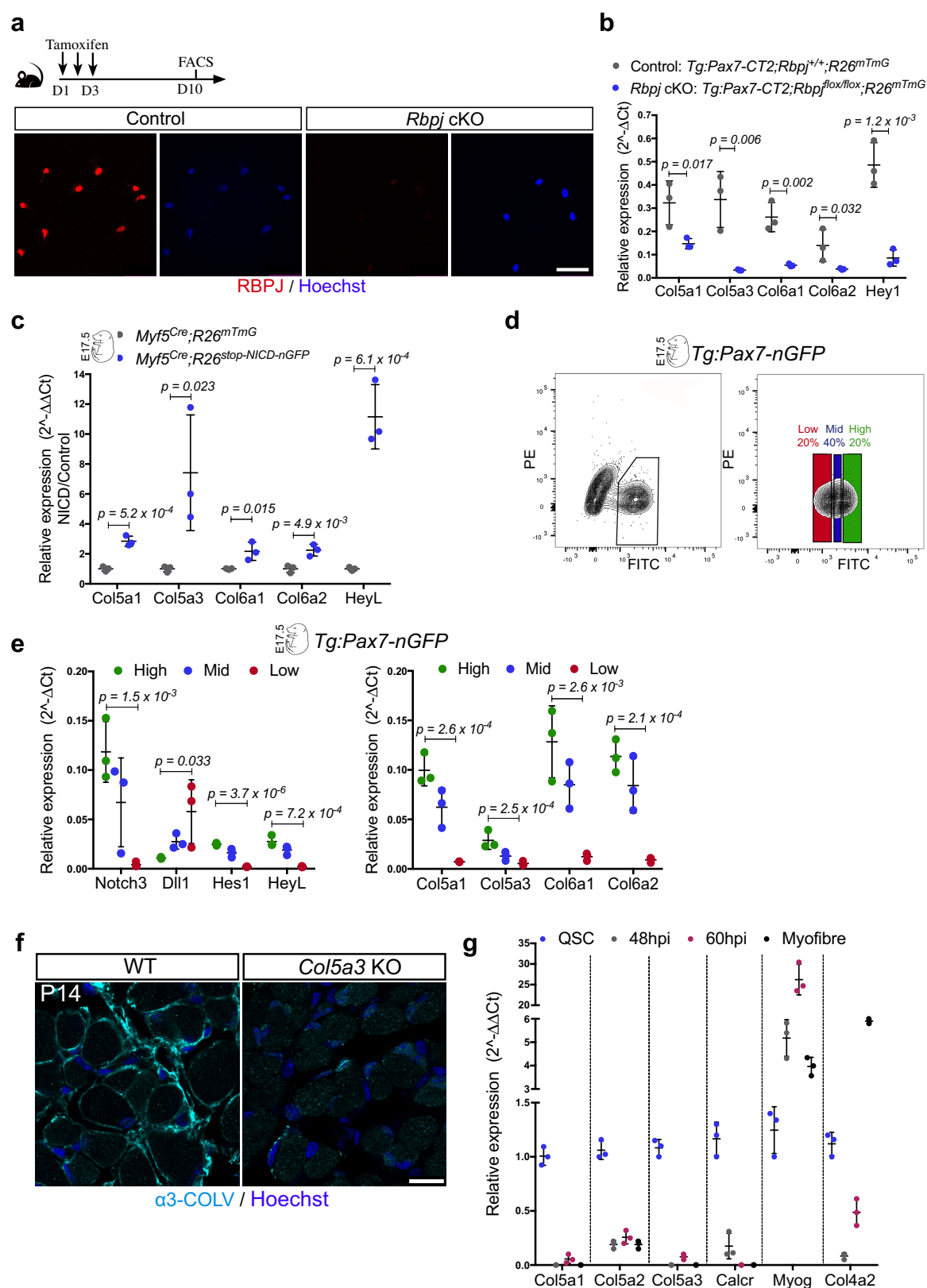
27. Haldar, M., Karan, G., Tvrdik, P. & Capocchi, M. R. Two cell lineages, *myf5* and *myf5*-independent, participate in mouse skeletal myogenesis. *Dev. Cell* **14**, 437–445 (2008).
28. Murphy, M. M., Lawson, J. A., Mathew, S. J., Hutcheson, D. A. & Kardon, G. Satellite cells, connective tissue fibroblasts and their interactions are crucial for muscle regeneration. *Development* **138**, 3625–3637 (2011).
29. Murtaugh, L. C., Stanger, B. Z., Kwan, K. M. & Melton, D. A. Notch signaling controls multiple steps of pancreatic differentiation. *Proc. Natl Acad. Sci. USA* **100**, 14920–14925 (2003).
30. Muzumdar, M. D., Tasic, B., Miyamichi, K., Li, L. & Luo, L. A global double-fluorescent Cre reporter mouse. *Genesis* **45**, 593–605 (2007).
31. Han, H. et al. Inducible gene knockout of transcription factor recombination signal binding protein-J reveals its essential role in T versus B lineage decision. *Int. Immunol.* **14**, 637–645 (2002).
32. Sun, M. et al. Collagen V is a dominant regulator of collagen fibrillogenesis: dysfunctional regulation of structure and function in a corneal-stroma-specific *Col5a1*-null mouse model. *J. Cell Sci.* **124**, 4096–4105 (2011).
33. Sambasivan, R. et al. Distinct regulatory cascades govern extraocular and pharyngeal arch muscle progenitor cell fates. *Dev. Cell* **16**, 810–821 (2009).
34. Hicks, C. et al. A secreted Delta1-Fc fusion protein functions both as an activator and inhibitor of Notch1 signaling. *J. Neurosci. Res.* **68**, 655–667 (2002).
35. Vasconcelos, F. F. et al. MyT1 counteracts the neural progenitor program to promote vertebrate neurogenesis. *Cell Reports* **17**, 469–483 (2016).
36. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_t}$ method. *Methods* **25**, 402–408 (2001).
37. Gao, M. et al. Discovery and optimization of 3-(2-(Pyrazolo[1,5-a]pyrimidin-6-yl)ethynyl)benzamides as novel selective and orally bioavailable discoidin domain receptor 1 (DDR1) inhibitors. *J. Med. Chem.* **56**, 3281–3295 (2013).
38. Shinin, V., Gayraud-Morel, B., Gomes, D. & Tajbakhsh, S. Asymmetric division and cosegregation of template DNA strands in adult muscle satellite cells. *Nat. Cell Biol.* **8**, 677–687 (2006).
39. Yoshida, N., Yoshida, S., Koishi, K., Masuda, K. & Nabeshima, Y. Cell heterogeneity upon myogenic differentiation: down-regulation of MyoD and Myf-5 generates 'reserve cells'. *J. Cell Sci.* **111**, 769–779 (1998).
40. Morita, S., Kojima, T. & Kitamura, T. Plat-E: an efficient and stable system for transient packaging of retroviruses. *Gene Ther.* **7**, 1063–1066 (2000).



Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Identification of NICD/RBPJ-bound enhancers and response to activation of Notch signalling. **a**, Gene expression microarray data show that satellite cells express a specific subset of collagen types, which include the fibrillar COLI (*Col1a1* and *Col1a2*), COLIII (*Col3a1*, possibly as $(\alpha 1(\text{III}))_3$ homodimer) and COLV (*Col5a1*, *Col5a2* and *Col5a3*) and the non-fibrillar COLIV (*Col4a1* and *Col4a2*), COLVI (*Col6a1* and *Col6a2*) and COLXV (*Col15a1*, possibly as $(\alpha 1(\text{XV}))_3$ homodimer). Data are shown as a heat map of normalized collagen transcripts expressed at different developmental time points (E12.5, E17.5 and post-natal day (P)8; *Tg-Pax7-nGFP*, Gene Expression Omnibus (GEO) accession number GSE52192), quiescent and post-injury ($t = 60$ h after BaCl_2 injury). **b**, ChIP-seq tracks indicating NICD/RBPJ-occupied enhancers, associated with mouse *Col5a1*, *Col5a3*, *Col6a1* and *Col6a2* loci. H3K4me1, H3K27ac, p300 and NICD are shown. Orange rectangles indicate RBPJ binding positions and asterisks indicate the enhancers used for transcriptional activity assays in **c**. **c**, Core sequences of the selected NICD/RBPJ-bound enhancers (asterisked orange rectangle in

Fig. 1a and in **b**). The RBPJ consensus binding motif is highlighted in yellow. **d**, Transcriptional response of isolated enhancers to activation of Notch signalling in C2C12 cells. Firefly luciferase signal was measured in cells with doxycycline-inducible expressed human Notch1–GFP (NICD) and GFP control cells treated with (2S)-N-[(3,5-difluorophenyl)acetyl]-L-alanyl-2-phenylglycine 1,1-dimethylethyl ester (DAPT) and were normalized to internal control (pCMV-*Renilla*). Data are expressed as relative luminescence units ($n = 3$ independent experiments). Data are mean \pm s.d.; two-sided paired *t*-test. **e**, Expression measurements, based on RNA sequencing, of collagen genes in myogenic C2C12 cells, with active (treated with Delta-like 1) or inhibited (treated with DAPT) Notch signalling for 6 or 24 h (data available at GEO, accession number GSE37184). Data are shown as Delta-like 1-to-DAPT ratios of average reads per kilobase of exon model per million mapped reads (RPKMs). Genes with low expression (RPKM < 2) were eliminated. *HeyL* and *Hey1* transcripts indicate Notch pathway activation. Red line designates no change (ratio = 1).

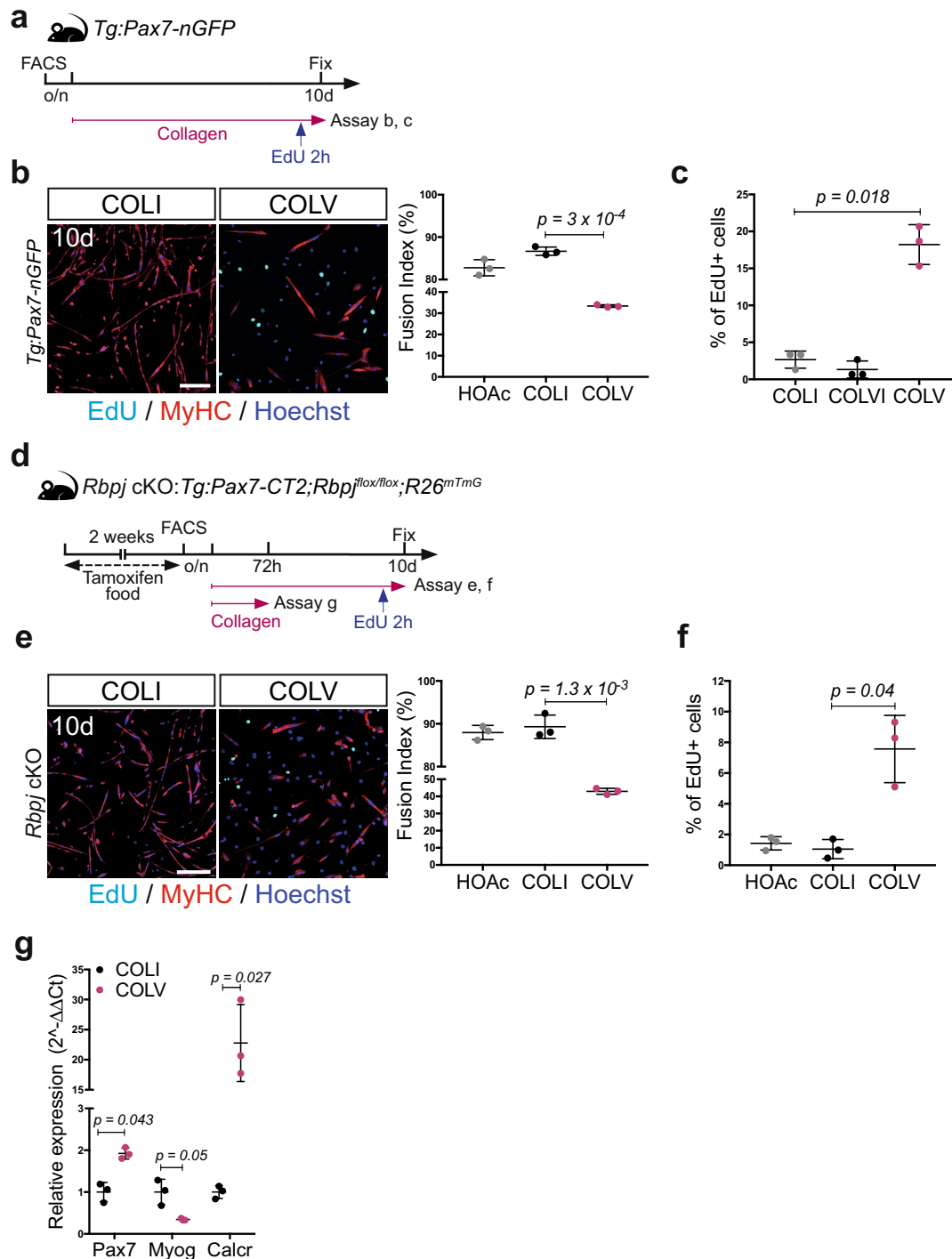


Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Notch signalling regulates *Col5* and *Col6* expression in vivo.

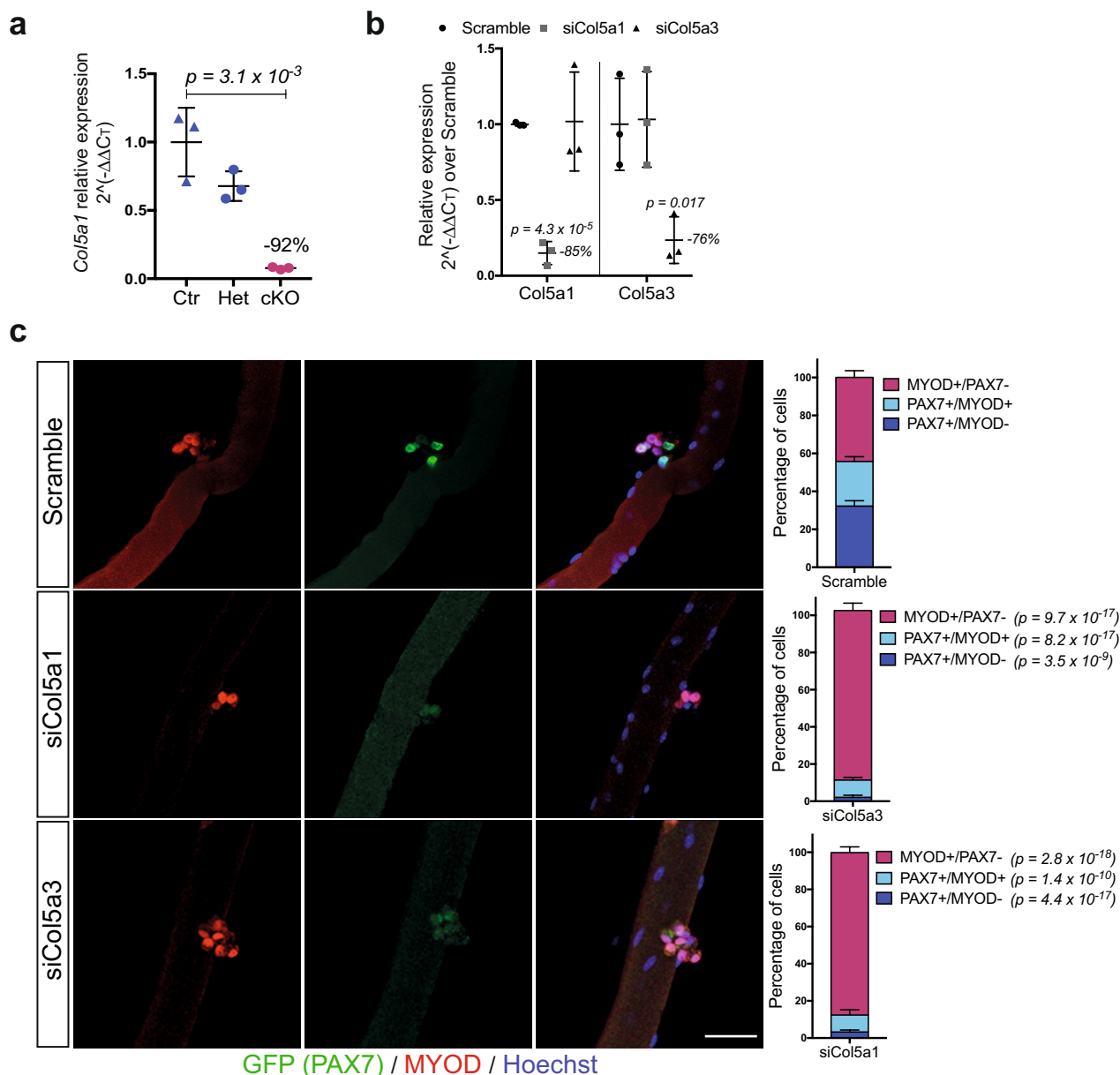
a, Satellite cells isolated by FACS at day 10 after tamoxifen injections, from resting tibialis anterior muscle from control (*Tg:Pax7-CT2;Rbpj*^{+/-}; *R26^{mTmG/+}*) and *Rbpj*-null (*Tg:Pax7-CT2;Rbpj*^{flax/-}; *R26^{mTmG/+}*) mice immunostained for RBPJ. **b**, RT-qPCR of collagen genes in *Rbpj* cKO and control satellite cells. *Hey1* used as control for Notch signalling ($n = 3$ mice per genotype). **c**, Induction of collagen genes in E17.5 control (*Myf5^{Cre/+};R26^{mTmG/+}*) and *Myf5^{Cre}-NICD* (*Myf5^{Cre/+};R26^{stop-NICD-nGFP/+}*) cells isolated by FACS. RT-qPCR was normalized to *Gapdh*, $n = 3$ fetuses per genotype. *HeyL* reports Notch activity. **d**, FACS plots showing fractionation of GFP⁺ cells from E17.5 *Tg:Pax7-nGFP* fetuses into Pax7^{high} (20% of population), Pax7^{mid} (40%), and Pax7^{low} (20%). The intensity of the GFP signal reflects the activity of the *Pax7* promoter. **e**, Transcript levels of GFP⁺ cells isolated by FACS show a tight

correlation between lineage progression, Notch signalling activity and collagen gene expression ($n = 3$ fetuses per genotype). **f**, Specificity of $\alpha 3$ -COLV antibody assessed by immunostaining of tibialis anterior muscle transverse section from wild-type and *Col5a3* cKO P14 postnatal pups ($n = 3$ mice per genotype). **g**, Time course of gene expression performed by RT-qPCR on freshly isolated satellite cells (Quiescent), 48 h or 60 h after cardiotoxin injury of tibialis anterior muscle (48 hours post injury (hpi), 60 hpi), and isolated single myofibres from extensor digitorum longus muscle of *Tg:Pax7-nGFP* mice. *Col5a1* and *Col5a3* were strongly downregulated in activated and differentiated cells. Quiescence (*Pax7*, *Calcr*) and differentiation (*Myog*) markers are indicated. *Col4a2*, a major component of the basement membrane, is expressed mainly by myofibres ($n = 3$ mice per condition). Data are mean \pm s.d.; one-sided unpaired *t*-test. Scale bars, 50 μ m.



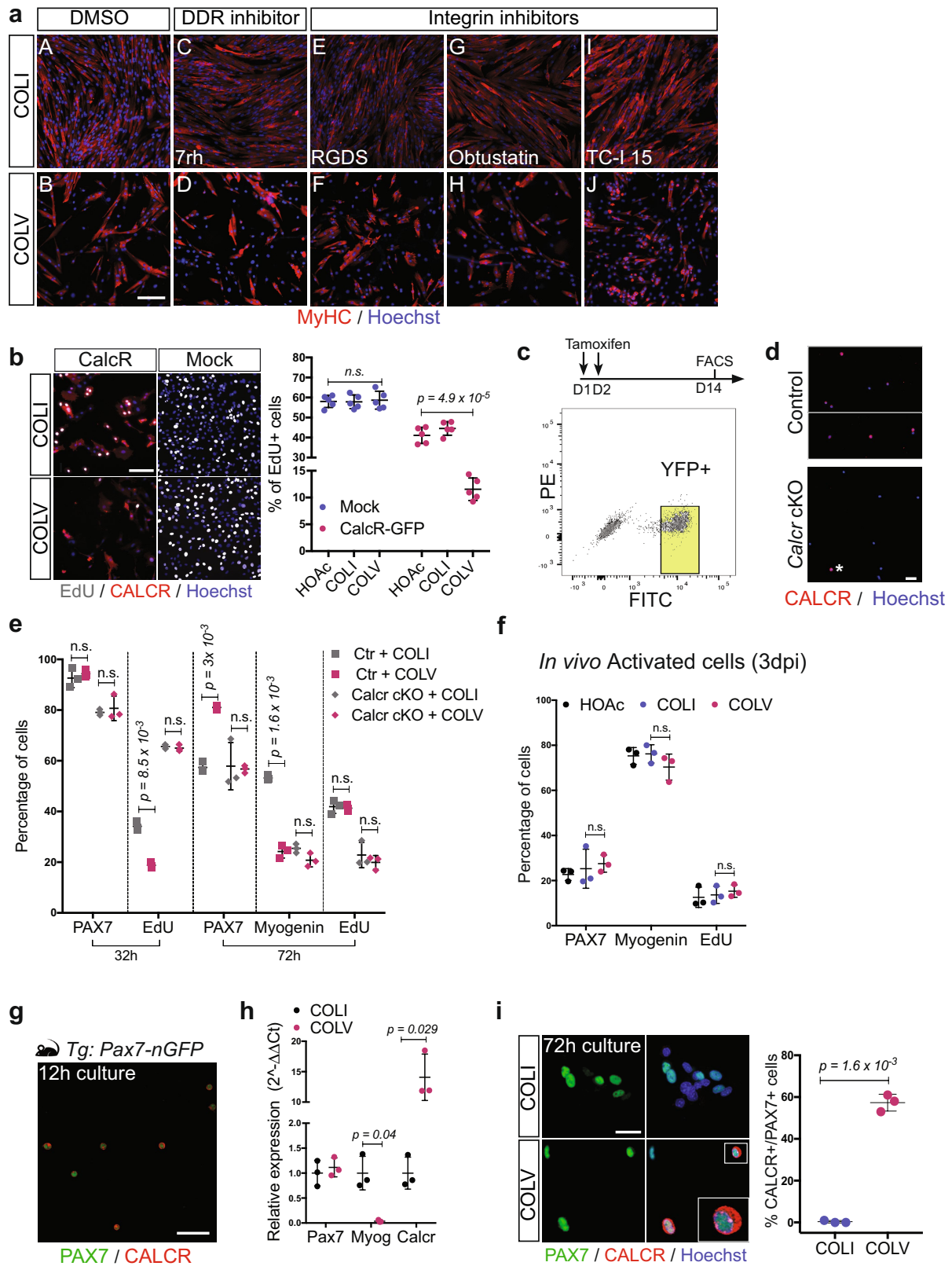
Extended Data Fig. 3 | COLV delays proliferation and differentiation of satellite cells. **a**, Experimental scheme: isolated *Tg:Pax7-nGFP* satellite cells cultured overnight (o/n) before collagen treatment. **b**, Myosin heavy chain (MyHC) and EdU staining of satellite cells treated with COLI or COLV. Fusion index: 82%, 86% and 33% for HOAc solvent, COLI and COLV, respectively ($n = 3$ mice, ≥ 250 cells, 2 wells per condition). **c**, Percentage of EdU⁺ primary myogenic cells after ten days of culture with indicated collagens. EdU: 2.6%, 1.3% and 18.2% for COLI, COLVI and COLV, respectively ($n = 3$ mice, ≥ 250 cells, 2 wells per condition). **d**, Experimental scheme for control and cKO mice. Satellite cells were plated overnight before collagen treatment. **e**, GFP and MyHC

immunostaining of *Rbpj* cKO satellite cells ($n = 3$ mice per condition) incubated 60 h in presence of COLI or COLV, or with HOAc control ($n = 3$ mice, ≥ 200 cells, 2 wells per condition). **f**, Percentage of EdU⁺ cells (2 h pulse) of *Rbpj*-null primary myogenic cells, after ten days of culture with HOAc or indicated collagens. EdU: 1.0% and 7.6% for COLI and COLV, respectively ($n = 3$ mice, ≥ 150 cells, 2 wells per condition). **g**, RT-qPCR on GFP⁺ *Rbpj*-null satellite cells isolated by FACS and cultured for 72 h in the presence of COLI or COLV. Results are normalized to *Tbp*. Data are mean \pm s.d.; two-sided paired *t*-test; *P* value: two-sided unpaired *t*-test. Scale bars, 50 μ m.



Extended Data Fig. 4 | COLV—and specifically $\alpha 3$ -COLV—is critical for satellite cell self-renewal. **a**, RT-qPCR of *Col5a1* in control (Ctrl; *Tg:Pax7-CT2;Col5a1*^{+/+}; *R26*^{mTmG}), heterozygous (Het; *Tg:Pax7-CT2;Col5a1*^{flox/+}; *R26*^{mTmG}) and conditional knockout (cKO; *Tg:Pax7-CT2;Col5a1*^{flox/flox}; *R26*^{mTmG}) mice two weeks after tamoxifen diet ($n = 3$ mice per genotype). **b**, Transcript levels of the different *Col5* mRNA chains in C2C12 after transfection of either control scramble, *Col5a1* or *Col5a3* siRNA, showing the specificity of each siRNA for its given targeted mRNA. Data are normalized to *Tbp* gene expression ($n = 3$ independent assays). **c**, *Col5a1* and *Col5a3* siRNA transfection

of *Tg:Pax7-nGFP* isolated single myofibres cultured for 72 h and immunostained for GFP and MYOD. Resident satellite cells enter the myogenic program and form clusters composed of proliferating (PAX7⁺MYOD⁺MYOG⁻), differentiated (PAX7⁻MYOG⁺) and self-renewed (PAX7⁺MYOD⁻) cells within 72 h. Quantification of PAX7⁺MYOD⁻, PAX7⁺MYOD⁺ and PAX7⁻MYOD⁺ populations 72 h after transfection. Scramble siRNA was used as negative control ($n \geq 15$ fibres counted from 3 mice). Data are mean \pm s.d.; **a**, two-sided unpaired *t*-test; **b**, **c**, two-sided paired *t*-test. Scale bar, 50 μ m.

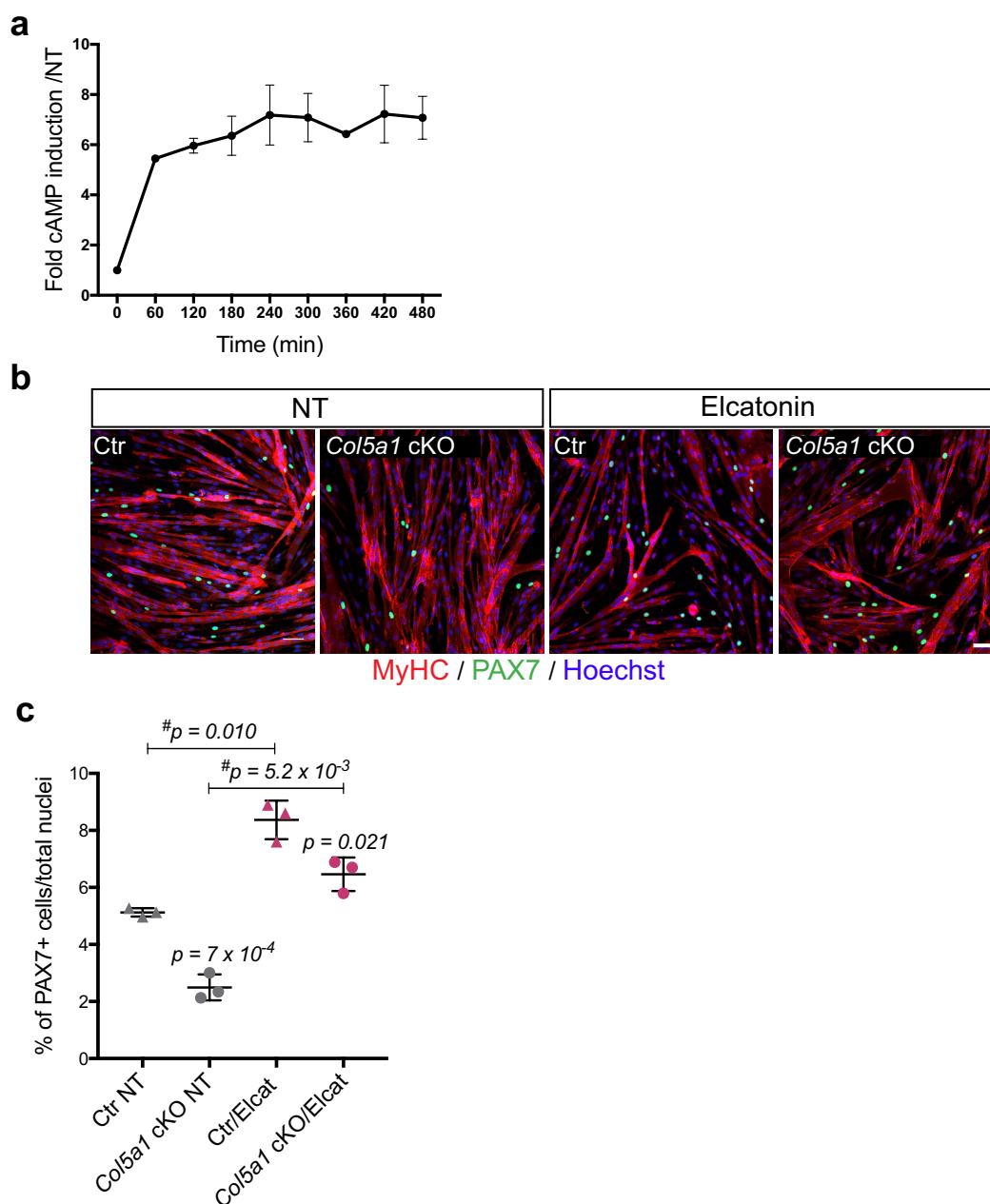


Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Screening for COLV receptor candidates

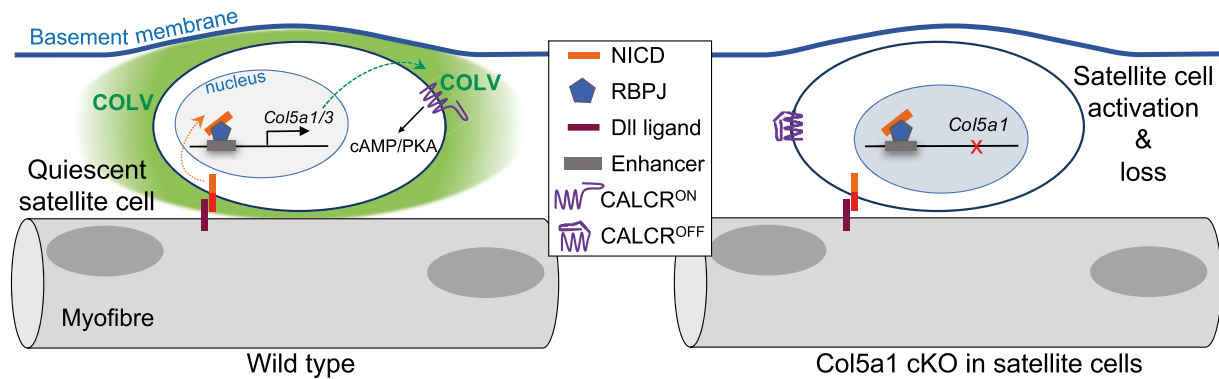
identifies CALCR. **a**, Screening for the COLV receptor: satellite cells from *Tg:Pax7-nGFP* mice were incubated for ten days with COLV and candidate receptors were targeted with respective inhibitors: 7rh for DDR1 (sub-panels C, D), the broad-spectrum Integrin-binding competitor RGDS peptide (sub-panels E, F), Obtustatin for Integrin $\alpha 1\beta 1$ (sub-panels G, H), TC-I 15 for Integrin $\alpha 2\beta 1$ (sub-panels I, J). DMSO solvent was used as a control for TC-I 15 and 7rh (sub-panels A, B). Satellite cell differentiation was assayed by MyHC immunostaining. **b**, EdU (2 h pulse) and CALCR staining of GFP⁺ C2C12 cells isolated by FACS and transduced with *Calcr*-GFP or mock-GFP retrovirus and cultured for 24 h with COLI (top) or COLV (bottom). Quantification of EdU⁺ *Calcr*-transduced C2C12 cells or mock-GFP cells treated for 24h with COLV or with the controls, COLI and HOAc ($n = 5$ independent experiments, ≥ 250 cells counted, 2 wells per condition). There was no significant difference between HOAc and COLI treated samples (data not shown). **c**, Experimental scheme of tamoxifen administration to control (Ctr) (*Calcr*^{+/+}) and cKO (*Calcr*^{flox/flox}) mice. FACS plot of satellite cells from *Pax7*^{CreERT2/+};*Calcr*^{flox/flox};*R26*^{stop-YFP} and *Pax7*^{CreERT2/+};*Calcr*^{+/+};*R26*^{stop-YFP} mice. Cells sorted based on YFP expression. **d**, Control and *Calcr* cKO satellite cells isolated by FACS, fixed immediately after sorting and immunostained for CALCR to confirm the absence of CALCR protein from recombined cells. For control (upper

panel), two fields from the same culture dish are shown, separated by a white line. Asterisk shows a non-recombined, CALCR⁺ cell in the cKO sample (lower panel). **e**, Quantification of PAX7⁺, Myogenin⁺ and EdU⁺ cells in *Calcr*-depleted satellite cells (*Pax7*^{CT2/+};*Calcr*^{flox/flox};*R26*^{stop-YFP}) isolated by FACS and treated for 32 h or 72 h with COLI or COLV ($n = 3$ mice, ≥ 250 cells counted, 2 wells per condition). **f**, Quantification of total PAX7⁺ (GFP), Myogenin⁺ and EdU⁺ myogenic cells isolated by FACS from *Tg:Pax7-nGFP* mice three days after cardiotoxin injury of tibialis anterior muscle, and incubated for 72 h in presence of COLI or COLV, or HOAc as a control, in the culture medium ($n = 3$ mice, ≥ 200 cells counted). **g**, CALCR protein in freshly isolated satellite cells, or satellite cells cultured for 12 h, from *Tg:Pax7-nGFP* mice, demonstrating that CALCR protein is still present when satellite cells are treated with different collagens (see Extended Data Fig. 2). **h**, Induction of *Calcr* transcript expression by RT-qPCR of *Tg:Pax7-nGFP* satellite cells isolated by FACS and cultured for 72 h in the presence of COLI or COLV. Results are normalized to *Tbp* ($n = 3$ mice). **i**, Immunostainings for CALCR protein of *Tg:Pax7-nGFP* satellite cells cultured for 72 h in presence of COLI or COLV ($n = 3$ mice, ≥ 50 cells, 2 wells per condition). Data are mean \pm s.d.; **b**, two-sided unpaired *t*-test; **c–i**, two-sided paired *t*-test. Scale bars, 25 μ m (**g**), 50 μ m (**a**, **b**, **d**, **i**).



Extended Data Fig. 6 | CALCR ligand elcatonin can substitute the depletion of the surrogate ligand COLV. a, Intracellular levels of cAMP in *Calcr*-transduced C2C12 cells treated with COLV for up to 480 min ($n = 4$ independent assays). **b,** Rescue of loss of COLV by elcatonin in an ex vivo self-renewal reserve-cell model, where PAX7⁺ non-proliferative cells return to quiescence (see Methods). MyHC and PAX7 staining of control (Ctr: *Tg:Pax7-CT2;Col5a1^{+/+};R26^{mTmG}*) and *Col5a1*-null

(*Tg:Pax7-CT2;Col5a1^{fllox/flox};R26^{mTmG}*) cells, non-treated (NT) or treated with elcatonin. No GFP⁺EdU⁺ cells (12 h pulse) could be detected under any of the conditions, indicating GFP⁺ cells are quiescent (data not shown). **c,** Quantification of percentage of reserve cells (PAX7⁺ per total nuclei) ($n = 3$ mice per genotype and condition, ≥ 350 cells counted). Elcat, elcatonin. Data are mean \pm s.d.; two-sided paired *t*-test; #, *P* value calculated by two-sided unpaired *t*-test. Scale bar, 50 μ m.



Extended Data Fig. 7 | Schematic of Notch–COLV–CALCR axis in satellite cells. A Notch–COLV–CALCR signalling cascade actively maintains satellite cell quiescence. Satellite cells are in direct contact with the plasma membrane of the myofibre (black outline) and an overlying basement membrane (blue line). Activation of the Notch receptor is achieved by a ligand (probably DLL1 or DLL4) present on the muscle fibre. Induction of *Col5a1* and *Col5a3* (and also *Col6a1* and *Col6a2*) genes occurs via distal regulatory elements (grey box). Satellite-cell-produced

COLV is deposited under the basement membrane and acts as a surrogate ligand of the plasma membrane receptor CALCR, also expressed by the satellite cells, thereby propagating a cell-autonomous signalling system in the local niche. In the absence of COLV (deletion of *Col5a1*) the quiescent niche is disturbed, CALCR signalling is abrogated, and satellite cells spontaneously differentiate and fuse to myofibres, leading to exhaustion of the muscle stem cell pool.

Reconstruction of antibody dynamics and infection histories to evaluate dengue risk

Henrik Salje^{1,2,3,4*}, Derek A. T. Cummings^{4,5,6}, Isabel Rodriguez-Barraquer⁷, Leah C. Katzelnick⁵, Justin Lessler⁴, Chonticha Klungthong⁸, Butsaya Thaisomboonsuk⁸, Ananda Nisalak⁸, Alden Weg⁸, Damon Ellison⁸, Louis Macareo⁸, In-Kyu Yoon⁹, Richard Jarman¹⁰, Stephen Thomas¹¹, Alan L. Rothman¹², Timothy Endy^{11,13} & Simon Cauchemez^{1,2,3,13}

As with many pathogens, most dengue infections are subclinical and therefore unobserved¹. Coupled with limited understanding of the dynamic behaviour of potential serological markers of infection, this observational problem has wide-ranging implications, including hampering our understanding of individual- and population-level correlates of infection and disease risk and how these change over time, between assay interpretations and with cohort design. Here we develop a framework that simultaneously characterizes antibody dynamics and identifies subclinical infections via Bayesian augmentation from detailed cohort data (3,451 individuals with blood draws every 91 days, 143,548 haemagglutination inhibition assay titre measurements)^{2,3}. We identify 1,149 infections (95% confidence interval, 1,135–1,163) that were not detected by active surveillance and estimate that 65% of infections are subclinical. After infection, individuals develop a stable set point antibody load after one year that places them within or outside a risk window. Individuals with pre-existing titres of $\leq 1:40$ develop haemorrhagic fever 7.4 (95% confidence interval, 2.5–8.2) times more often than naive individuals compared to 0.0 times for individuals with titres $> 1:40$ (95% confidence interval: 0.0–1.3). Plaque reduction neutralization test titres $\leq 1:100$ were similarly associated with severe disease. Across the population, variability in the size of epidemics results in large-scale temporal changes in infection and disease risk that correlate poorly with age.

Despite the large body of literature from observational and cohort studies describing dengue cases, we still have major difficulties in explaining individual- and population-level differences in infection and disease risk. These difficulties mostly arise from a fundamental methodological issue in the research of many pathogens for which individual histories of infection are difficult to capture. The four dengue virus serotypes (DENV1–DENV4), which are found across tropical and sub-tropical regions and lead to an estimated 390 million infections each year, cause a range of disease manifestations, from asymptomatic infection to death^{4,5}. High levels of subclinical infection indicate that even in regions with thorough active surveillance, the majority of infections are missed¹. This observational problem has wide ranging implications as it not only hampers our ability to estimate the underlying level of infection in the community and characterize individual risk factors for infection and severity, but also our ability to assess correlates of protection, dynamically monitor susceptibility at both the population and individual level, define optimal thresholds for the interpretation of serological assays and critically assess cohort design.

Here, we develop an analytical framework that can address this challenge, leading to new insights into a broad range of questions. We use this framework to both characterize antibody changes following

infection and identify infection events that were missed by surveillance on the basis of the analysis of longitudinal data from cohort studies. We apply the analysis to data from a school-based cohort study in Thailand ($n = 3,451$, mean age at recruitment was 9 years old, inter-quartile range, 8–11), in which subjects had blood taken on average every 91 days for up to five years and when illnesses were detected through active surveillance². Surveillance of active fever and school absence of children was conducted from June to mid-November when DENV circulation is concentrated². Haemagglutination inhibition tests were used to measure antibody titres of each serotype in each sample (143,548 haemagglutination inhibition measurements in total). Plaque reduction neutralization test (PRNT) titres were also measured on a subset of 1,771 samples. Haemagglutination inhibition titres correlated closely with PRNT titres (Pearson correlation of 0.91) and with inhibition enzyme-linked immunosorbent assays (ELISAs), although titre values differ between laboratories and between assays^{6–9}.

To track the evolution in the measured antibody titres of an individual (Fig. 1a), we placed titres on an adjusted \log_2 scale (titres of 1:10 were given a value of 1, 1:20 a value of 2 and so on). There were 274 detected symptomatic DENV infections (Fig. 1b); 62 children were hospitalized (23%) and 36 had dengue haemorrhagic fever (DHF) (13%). In cases for which the infecting serotype was known through PCR (79% of cases, Supplementary Table 1), we observed a sharp rise and subsequent decay in \log_2 titres after the onset of symptoms (Fig. 1c, d). The mean \log_2 titre of the infecting serotype was 0.79 (95% confidence interval, 0.74–0.84) times the \log_2 titre of the non-infecting serotype in the three months before onset of symptoms compared to 0.94 (95% confidence interval, 0.93–0.96) times in the six months after the onset of symptoms (Fig. 1e). Because 86% of cases with symptomatic infections had detectable titres of at least one serotype before infection, the higher antibody titre of non-infecting serotypes probably captures responses to prior infections¹⁰.

We reconstructed the antibody trajectories of each individual by assuming that infection leads to an increase in titres that subsequently decays exponentially¹¹. We also explored biphasic responses (Extended Data Fig. 1). We allow for variability in antibody kinetics across individuals and infections, and for differential rises for the infecting versus the non-infecting serotypes for primary infections but undifferentiated responses for subsequent infections. We use data augmentation techniques to impute undetected infections (subclinical infections during active surveillance or unknown symptom status outside the surveillance windows) and to identify the serotype of undetected primary infections³. Instead of relying on fixed cut-offs to identify infections, data augmentation allows us to incorporate uncertainty in the existence, timing and serotype of unobserved infection events and therefore

¹Mathematical Modelling of Infectious Diseases Unit, Institut Pasteur, Paris, France. ²CNRS UMR2000, Génomique évolutive, modélisation et santé (GEMS), Institut Pasteur, Paris, France. ³Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, Paris, France. ⁴Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA. ⁵Department of Biology, University of Florida, Gainesville, FL, USA. ⁶Emerging Pathogens Institute, University of Florida, Gainesville, FL, USA. ⁷University of California, San Francisco, San Francisco, CA, USA. ⁸Department of Virology, Armed Forces Research Institute of Medical Sciences, Bangkok, Thailand. ⁹International Vaccine Institute, Seoul, South Korea. ¹⁰Viral Diseases Branch, Walter Reed Army Institute of Research, Silver Spring, MD, USA. ¹¹Department of Medicine, Upstate Medical University of New York, Syracuse, NY, USA. ¹²Institute for Immunology and Informatics, Department of Cell and Molecular Biology, University of Rhode Island, Providence, RI, USA. ¹³These authors jointly supervised this work: Timothy Endy, Simon Cauchemez.

*e-mail: henrik.salje@pasteur.fr

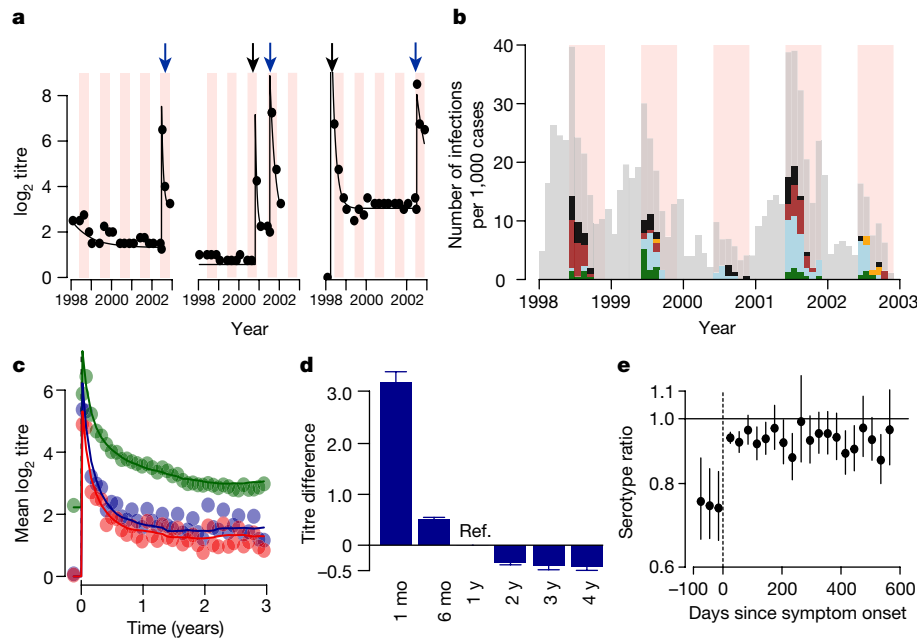


Fig. 1 | Titre responses following infection. **a**, Measured (dots) and model fit (lines) for three example individuals. Each dot represents the mean titre across the four serotypes. The pink-shaded regions are periods of active surveillance. Blue arrows represent confirmed symptomatic dengue infections. Black arrows represent estimates of timing of subclinical infections from an augmented dataset. During the active surveillance windows, these augmented infections represent subclinical infections whereas outside the surveillance window, it is unknown if the individual had symptoms. **b**, Serotype distribution of PCR-confirmed symptomatic infections (green, DENV1; blue, DENV2; maroon, DENV3; orange, DENV4; black, unknown serotype). The grey bars represent the estimated distribution of infections not detected by active surveillance.

we can probabilistically assess whether differences in measured titres are due to infections or assay variability.

We find that after post-primary infections, there is a mean increase of 5.8 (95% confidence interval, 5.6–5.9) in \log_2 titres across serotypes, which decreases by 76% after one year. For primary infections (that is, individuals without detectable titres before infection), the mean increase in \log_2 titre is 7.6 (95% confidence interval, 7.4–7.8) for the infecting serotype and 6.6 for non-infecting serotypes (95% confidence interval, 6.4–6.7). The similarity in titres of infecting and non-infecting serotypes coupled with assay variability suggests that in a clinical setting individual haemagglutination inhibition measurements cannot reliably determine the infecting serotype. We find that titres largely stabilize one year after infection to a set point (the ‘set-point antibody load’; Fig. 1d). There is substantial variability between infections: the interquartile range of the increase in \log_2 titre one year after infection is 0.7–2.2 across all infections (Extended Data Fig. 2a). We find that even after accounting for historic infection status, measured DENV2 titres are systematically lower than other serotypes (0.85 lower than DENV1; Extended Data Fig. 2b and Supplementary Table 2), which could indicate technical considerations of the DENV2 assay or inherent differences in immune responses to DENV2. We estimate the measurement error in the haemagglutination inhibition assay (that is, the standard deviation in any reading) to be 0.49 (95% confidence interval, 0.49–0.50), which is similar to the empirically estimated standard deviation using repeated testing on the same serum and 2.6 times the error estimates of the PRNT¹² (Extended Data Fig. 2c). Despite the variability in individual readings, because we use many readings from four serotypes for each participant and titres appear to behave in a stable and predictable manner, we can nevertheless make robust inferences when considering the ensemble of the measurements.

We probabilistically identify 1,149 undetected infections (95% range across model iterations, 1,135–1,163), of which 507 (494–520) occurred

during active surveillance periods and were therefore subclinical (Fig. 1b). Overall, we estimate that 35% of infections are symptomatic (95% confidence interval, 34–36%). The temporal distribution of subclinical infections was correlated with that of symptomatic infections (Pearson correlation of 0.78, 95% confidence interval, 0.70–0.84). Using augmented primary infections for cases in which we could confidently assign the infecting serotype (same serotype implicated by >50% of iterations), we find that 34% of undetected primary infections (and 39% of subclinical primary infections) were due to DENV4, compared to only 3% of all symptomatic infections (none of which were primary infections; Extended Data Fig. 3a, b). We find consistent results using a more stringent cutoff to assign the infecting serotype (Extended Data Fig. 3c). These findings are consistent with a reduced risk of disease from DENV4 compared to other serotypes resulting in a mostly silent DENV4 epidemic. This is supported by a phylogenetic analysis that found that DENV4 was widespread in Thailand throughout this period (see supplementary figure 4 of Salje et al.¹³). This suggests that the serotype distributions from hospital-based or community-based surveillance may not be representative of infections in the population and supports previous evidence that the transmissibility of a serotype can be delinked from the propensity to cause symptomatic and/or severe disease^{14,15}. Furthermore, these results indicate that factors that contribute to transmission potential (for example, viral replication, peak titres or infection length) are not predictive of adverse outcomes¹⁶.

We find that the underlying probability of infection and the probability of developing disease are strongly linked to the mean antibody titre at the time of exposure. Overall, an individual’s annual risk of infection was 17%, varying from 21% for individuals with mean measured \log_2 titres of <2, to 16% for those with \log_2 titres of 2–3 and 11% for those with \log_2 titres of >3 (Fig. 2a). Using logistic regression, we find that for \log_2 titres of >2, each unit increase in \log_2 titres is associated with a $0.71 \times$ relative risk of infection (95% confidence interval, 0.67–0.76).

We find that the underlying probability of infection and the probability of developing disease are strongly linked to the mean antibody titre at the time of exposure. Overall, an individual’s annual risk of infection was 17%, varying from 21% for individuals with mean measured \log_2 titres of <2, to 16% for those with \log_2 titres of 2–3 and 11% for those with \log_2 titres of >3 (Fig. 2a). Using logistic regression, we find that for \log_2 titres of >2, each unit increase in \log_2 titres is associated with a $0.71 \times$ relative risk of infection (95% confidence interval, 0.67–0.76).

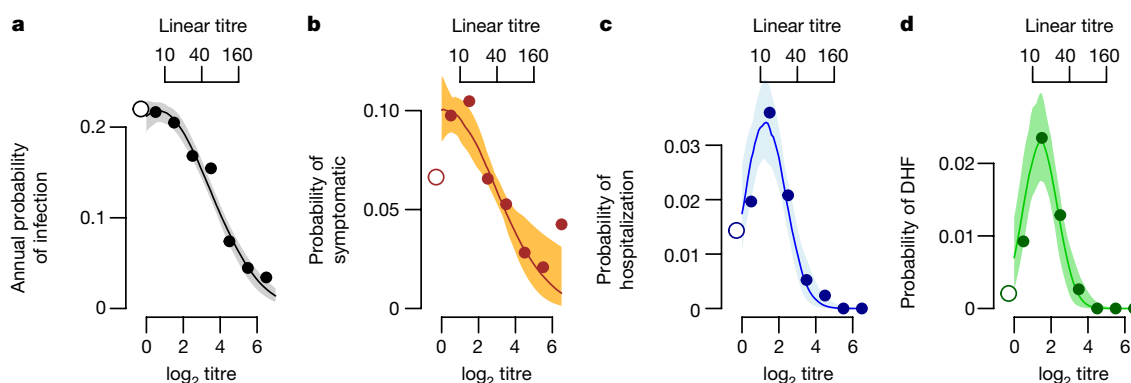


Fig. 2 | Probability of infection and disease as a function of titre. **a–d**, Annualized probability of infection (**a**), developing any symptoms (**b**), being hospitalized (**c**) and developing DHF (**d**) as a function of the mean measured antibody titre across all serotypes at the time of exposure

The annual probability of having a symptomatic infection varies from 6.4% (95% confidence interval, 4.9–8.4%) for primary infections to 8.4% (95% confidence interval, 7.8–9.1%) for individuals with pre-existing \log_2 titres ≤ 3 ($\leq 1:40$ on a linear scale) and 4.0% (95% confidence interval, 3.0–5.0%) for those with \log_2 titres > 3 (Fig. 2b). The annual probability of being hospitalized during a primary infection was 1.2% (95% confidence interval, 0.5–2.1%), compared to 2.4% (95% confidence interval, 2.1–2.7%) during a subsequent infection for those with pre-existing \log_2 titres ≤ 3 and 0.3% for those with \log_2 titres > 3 (95% confidence interval, 0.09–0.6%; Fig. 2c). Even more pronounced was the risk for developing DHF, which ranged from 0.2% (95% confidence interval, 0.0–0.6%) for primary infections to 1.5% (95% confidence interval, 1.3–1.7%) for subsequent infections in those patients with \log_2 titres ≤ 3 and 0.0% for \log_2 titres > 3 (95% confidence interval, 0.0–0.4%; Fig. 2d). Within this study population, an average of 54% of the population had detectable \log_2 titres of ≤ 3 at any point in time. Time-varying Cox proportional hazards models that specifically account for the dependence of titre observations within individuals gave similar results¹⁷ (Extended Data Fig. 4). Using \log_2 titres to probabilistically identify the cohort participants with detectable titres that will develop DHF has an area under the curve (AUC) value of 0.66 (Extended Data Fig. 5).

When considering only infected individuals, we observe no difference in the probability of subclinical infection by titre; however, the probability of hospitalization and DHF remains the highest in those with pre-existing \log_2 titres of ≤ 3 (Extended Data Fig. 6a–c). Only one individual with pre-infection \log_2 titres > 3 developed DHF during surveillance compared to 146 who did not, but who had titres at infection within the same range. This suggests that in the event that infection does take place, antibodies are not protective against developing symptoms as such, but conversely, are associated with the development of severe disease. We observe no difference in the risk of disease given infection across years (Supplementary Table 3) or age (Supplementary Table 4). Other studies are needed to investigate whether younger age groups than those included here nevertheless have an increased risk. PRNTs form the basis of current discussions on immune correlates. Among those infected, individuals with detectable PRNT \log_2 titres of ≤ 4.5 (equivalent to approximately $\leq 1:100$) have a 7.5 times (95% confidence interval, 2.4–11.6) higher risk of DHF compared to previously naive individuals, compared to 0.0 times for those with higher titres (Extended Data Fig. 6d–f). Cross-reactive titres that result from exposure to non-DENV flaviviruses such as Japanese encephalitis and Zika may be included in these risk estimates.

Our findings suggest that after infection set-point antibody loads appear to be important for the determination of individual infection and disease risk. After infection, we estimate the daily probability of a subsequent infection and the development of DHF disease as a function

across all study subjects ($n = 3,451$). The open circles on the left represent primary infections (that is, those with no detectable titres for any serotype before exposure). The shaded regions represent 95% bootstrap confidence intervals.

of titre dynamics. We demonstrate that the probability of both infection and disease stabilizes after one year (Fig. 3). On the basis of the observation in Fig. 2 that individuals with detectable titres of ≤ 3 had an increased risk of infection and disease, we explored the temporal evolution of risk following infection for those with set-point antibody loads (that is, the titre at one year following infection) above and below this threshold. At one year, we observe a 2.1 times higher risk of infection (irrespective of disease outcome) for those with set-point antibody loads of ≤ 3 compared to those with greater antibody loads and an 8.9 times higher risk of infection that leads to DHF. Overall, we find that three years after infection, 34% of individuals with set-point antibody loads of ≤ 3 suffer a subsequent infection, irrespective of severity (95% confidence interval, 33–35%) compared to 23% for those with greater loads (95% confidence interval, 20–26%). After this delay, 3.5% of individuals with set-point loads of ≤ 3 develop DHF disease (2.4–4.4%) compared to none in those with higher loads. The apparent stability of set-point antibody loads points to an ability to assess the long-term risk of an individual.

Our findings are consistent with low titres generated by some candidate vaccines in previously naive individuals, ‘priming’ individuals for severe disease upon their first exposure¹⁸. A hypothesis that is supported by previous evidence that primary infections in infants with maternal antibodies and secondary infections in older individuals are associated with severe disease^{19,20}. Furthermore, a Nicaraguan study found elevated risk of severe disease for those with low inhibition ELISA titres at prior annual blood draws⁹. Previously naive individuals given the dengvaxia vaccine had mean PRNT titres within our risk window²¹ (Fig. 4d). Further work is required to understand whether immunity acquired from vaccination and natural infection are qualitatively similar and whether the risk window described here is relevant for vaccine recipients. T cell immunity, which is not captured by these assays, might compensate for antibody titres in this window. Vaccine studies should carefully assess the criteria used to define seroconversion, and how titres correlate with disease risk over time. Our work suggests that previously used criteria (PRNT titre $> 1:10$) do not adequately correlate with reduction in disease risk and suggest that haemagglutination inhibition titres $> 1:40$ or PRNT titres of $> 1:100$ may provide a starting point for any vaccine in identifying a targeted neutralizing antibody response. Placebo data from the dengvaxia vaccine trials also suggests higher PRNT titres are linked to protection²². The targeted vaccination of individuals that have pre-existing antibody titres within our zone may be a viable approach to minimize the public health burden from dengue by moving individuals away from the risk window (Fig. 4d). Even in an endemic setting such as our cohort, there is considerable temporal variability in the serological status of 9-year-old individuals (Extended Data Fig. 7), suggesting that the current WHO guidance surrounding dengvaxia or similar guidance that

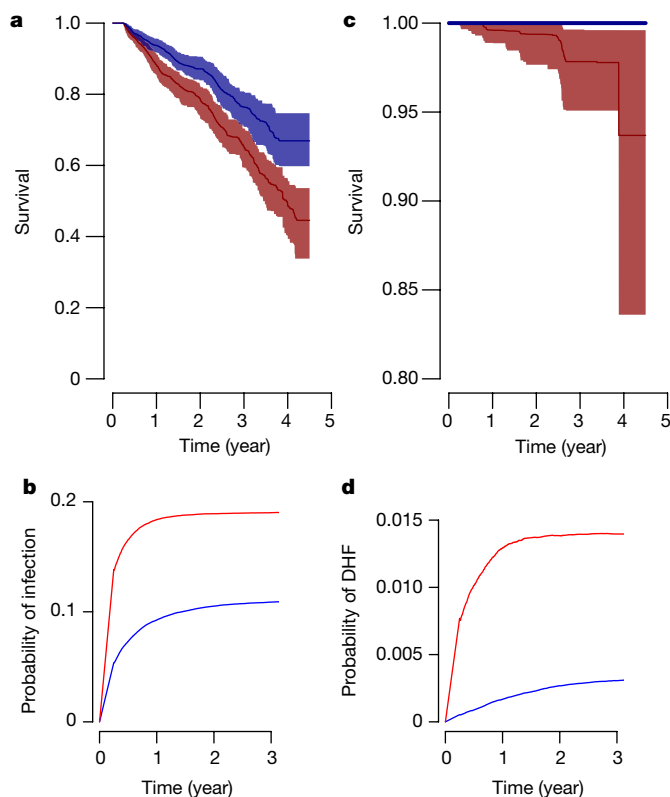


Fig. 3 | Risk of subsequent infection and disease following an infection event. Data are from the average of 1,420 infections across 100 reconstructed datasets. **a, c,** The probability of survival from subsequent infection (irrespective of disease outcome (**a**) and for those that led to DHF (**c**)) as calculated from Kaplan–Meier curves for those infections with set-point antibody titres of ≤ 3 (red) and > 3 (blue) with 95% confidence intervals. The annualized probability of a subsequent infection (irrespective of disease outcome (**b**) and for those that led to DHF (**d**)) at different time points following infection for those infections with set-point antibody titres of ≤ 3 (red) and > 3 (blue).

is based on serostatus at vaccination will have to carefully consider this variation or specifically screen individuals.

Our approach allows us to consider wider problems concerning drivers of dengue epidemiology. The assumption that population-wide immunity varies across time and dictates multi-annual dynamics of dengue pervades the literature and dominates current hypotheses about what drives large outbreaks of dengue in particular settings^{18,23–26}. More generally, the idea that temporally varying population immunity drives temporal dynamics of pathogens pervades infectious disease epidemiology^{27–29}. However, quantitative evidence that any population varies in dengue immune status over time is mostly unavailable, as is a link between the immune status of a population and the risk of epidemics in empirical data. Here, although we have only a short time series, we show that underlying the heterogeneity in the size of annual epidemics indicates that the risk of having titres within the risk zone for different birth cohorts is more correlated with the epidemic time point (Fig. 4a, mean correlation of 0.70) than with age (Fig. 4b, mean correlation of 0.23). Although both the probability of being naive and having \log_2 titres above the risk zone correlated with age, strong birth-cohort effects also exist (Extended Data Fig. 7). For example, among 9-year olds, we observe up to a twofold difference in the probability of being naive, depending on the year of the study.

Finally, our results can guide the design of cohort studies aiming to characterize transmission. Studies typically use a fourfold rise in titres against any serotype as evidence of infection, regardless of the timing of sample collection. Using our titre trajectories, we find that if blood draws are every 90 days, a fourfold cutoff point on measured titres has a specificity of $>99\%$ and a sensitivity of 87% (Fig. 4c and

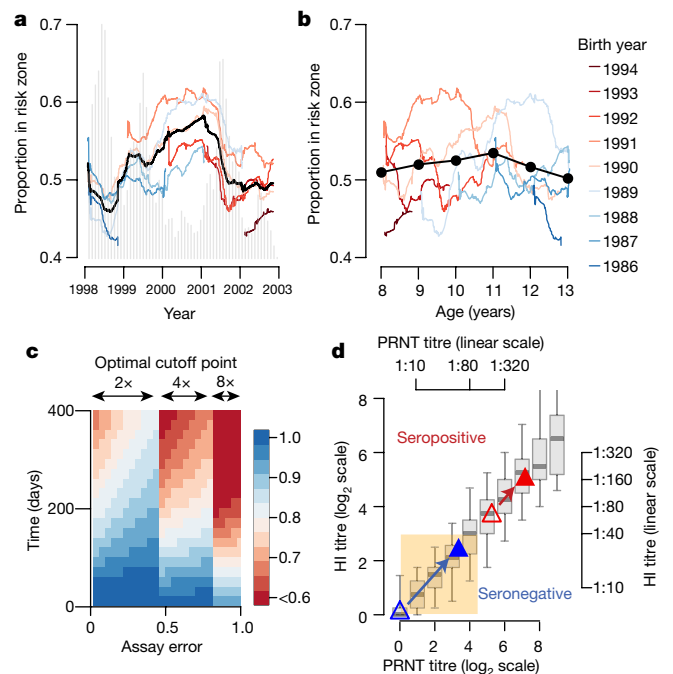


Fig. 4 | Evolution of population risk, implications for vaccine and cohort design. **a,** Proportion of study participants who have titres in the risk zone (defined as detectable \log_2 titres ≤ 3) during the study period for different birth cohorts (coloured lines) and overall (black). The epidemic curve of all infections is shown in grey. **b,** Proportion of study participants with titres in the risk zone as a function of age for different birth cohorts (coloured lines) and overall (black). **c,** Performance of current assay testing protocol for which infection events are defined as a rise above a cutoff point in any serotype across two blood draws. **d,** Relationship between PRNT titre and haemagglutination inhibition (HI) titre for samples for which both assays were performed ($n = 1,771$ samples). The box plots show 2.5, 25, 75 and 97.5 quantiles as well as the mean. Superimposed are the results from the denxavia vaccine study²¹ for previously seronegative (blue) and seropositive (red) samples before (open symbols) and after (filled symbols) vaccination.

Extended Data Fig. 8). The sensitivity is reduced to 77% when blood is taken every six months and 62% when blood is taken annually, although it may be higher in seasonal settings when samples are taken at the end of the season. Using an alternative approach that uses the mean titre across the four serotypes and a 1.6-fold cutoff point, the sensitivity of the assay improves to 96% when samples are taken every six months and to 90% for annual bleeds (specificity $>95\%$; Extended Data Fig. 9). We provide the optimal cutoff point and estimated sensitivity for these approaches and a theoretical estimate in which titres are on a continuous scale (such as PRNT) and for which a minimum specificity of $>99\%$ is required (Extended Data Fig. 9).

We demonstrate through simulation that our framework can recover the true number of subclinical infections and parameters when only 30% of infections are symptomatic (Supplementary Table 5). Our approach is also robust to a scenario in which there are differential rises in titres for symptomatic and non-symptomatic infections (Supplementary Table 6) and in which we incorporate school-specific force of infection parameters (Supplementary Table 7). In addition, we find that the timing (Extended Data Fig. 10a) and the serotype (Extended Data Fig. 10b) of undetected infections cluster in the same locations as symptomatic infections. This provides strong support for our modelling framework by suggesting that the model can correctly identify spatiotemporal clustering of otherwise undetected infections. These findings also support focal transmission, irrespective of disease outcome^{13,30,31}. The approach presented here will be applicable across disease systems for which longitudinal titre data exists, allowing a wide range of insights into fundamental questions of disease ecology and risk.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0157-4>.

Received: 18 December 2017; Accepted: 24 April 2018;

Published online 23 May 2018.

- Undurraga, E. A., Halasa, Y. A. & Shepard, D. S. Use of expansion factors to estimate the burden of dengue in Southeast Asia: a systematic analysis. *PLoS Negl. Trop. Dis.* **7**, e2056 (2013).
- Endy, T. P. et al. Epidemiology of inapparent and symptomatic acute dengue virus infection: a prospective study of primary school children in Kamphaeng Phet, Thailand. *Am. J. Epidemiol.* **156**, 40–51 (2002).
- Cauchemez, S. & Ferguson, N. M. Methods to infer transmission risk factors in complex outbreak data. *J. R. Soc. Interface* **9**, 456–469 (2012).
- Bhatt, S. et al. The global distribution and burden of dengue. *Nature* **496**, 504–507 (2013).
- Halstead, S. B. *Dengue* (Imperial College Press, London, 2008).
- Vaughn, D. W. et al. Dengue in the early febrile phase: viremia and antibody responses. *J. Infect. Dis.* **176**, 322–330 (1997).
- Harris, E. et al. Clinical, epidemiologic, and virologic features of dengue in the 1998 epidemic in Nicaragua. *Am. J. Trop. Med. Hyg.* **63**, 5–11 (2000).
- Venturi, G. et al. Humoral immunity and correlation between ELISA, hemagglutination inhibition, and neutralization tests after vaccination against tick-borne encephalitis virus in children. *J. Virol. Methods* **134**, 136–139 (2006).
- Katzelnick, L. C. et al. Antibody-dependent enhancement of severe dengue disease in humans. *Science* **358**, 929–932 (2017).
- Halstead, S. B., Rojanasuphot, S. & Sangkawibha, N. Original antigenic sin in dengue. *Am. J. Trop. Med. Hyg.* **32**, 154–156 (1983).
- Clapham, H. E. et al. Dengue virus (DENV) neutralizing antibody kinetics in children after symptomatic primary and postprimary DENV infection. *J. Infect. Dis.* **213**, 1428–1435 (2016).
- Salje, H. et al. Variability in dengue titer estimates from plaque reduction neutralization tests poses a challenge to epidemiological studies and vaccine development. *PLoS Negl. Trop. Dis.* **8**, e2952 (2014).
- Salje, H. et al. Dengue diversity across spatial and temporal scales: local structure and the effect of host population size. *Science* **355**, 1302–1306 (2017).
- Rodríguez-Barraquer, I. et al. Revisiting Rayong: shifting seroprofiles of dengue in Thailand and their implications for transmission and control. *Am. J. Epidemiol.* **179**, 353–360 (2014).
- Fried, J. R. et al. Serotype-specific differences in the risk of dengue hemorrhagic fever: an analysis of data collected in Bangkok, Thailand from 1994 to 2006. *PLoS Negl. Trop. Dis.* **4**, e617 (2010).
- Duong, V. et al. Asymptomatic humans transmit dengue virus to mosquitoes. *Proc. Natl Acad. Sci. USA* **112**, 14688–14693 (2015).
- D'Agostino, R. B. et al. Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham Heart Study. *Stat. Med.* **9**, 1501–1515 (1990).
- Ferguson, N. M. et al. Benefits and risks of the Sanofi–Pasteur dengue vaccine: modeling optimal deployment. *Science* **353**, 1033–1036 (2016).
- Kliks, S. C., Nimmanitya, S., Nisalak, A. & Burke, D. S. Evidence that maternal dengue antibodies are important in the development of dengue hemorrhagic fever in infants. *Am. J. Trop. Med. Hyg.* **38**, 411–419 (1988).
- Guzmán, M. G., Alvarez, M. & Halstead, S. B. Secondary infection as a risk factor for dengue hemorrhagic fever/dengue shock syndrome: an historical perspective and role of antibody-dependent enhancement of infection. *Arch. Virol.* **158**, 1445–1459 (2013).
- Villar, L. et al. Efficacy of a tetravalent dengue vaccine in children in Latin America. *N. Engl. J. Med.* **372**, 113–123 (2015).
- Moodie, Z. et al. Neutralizing antibody correlates analysis of tetravalent dengue vaccine efficacy trials in Asia and Latin America. *J. Infect. Dis.* **217**, 742–753 (2018).
- Cummings, D. A. T., Schwartz, I. B., Billings, L., Shaw, L. B. & Burke, D. S. Dynamic effects of antibody-dependent enhancement on the fitness of viruses. *Proc. Natl Acad. Sci. USA* **102**, 15259–15264 (2005).
- Wearing, H. J. & Rohani, P. Ecological and immunological determinants of dengue epidemics. *Proc. Natl Acad. Sci. USA* **103**, 11802–11807 (2006).
- Flasche, S. et al. The long-term safety, public health impact, and cost-effectiveness of routine vaccination with a recombinant, live-attenuated dengue vaccine (dengvaxia): a model comparison study. *PLoS Med.* **13**, e1002181 (2016).
- Adams, B. et al. Cross-protective immunity can account for the alternating epidemic pattern of dengue virus serotypes circulating in Bangkok. *Proc. Natl Acad. Sci. USA* **103**, 14234–14239 (2006).
- Grenfell, B. T., Bjørnstad, O. N. & Kappey, J. Travelling waves and spatial hierarchies in measles epidemics. *Nature* **414**, 716–723 (2001).
- Earn, D. J., Rohani, P., Bolker, B. M. & Grenfell, B. T. A simple model for complex dynamical transitions in epidemics. *Science* **287**, 667–670 (2000).
- Cobey, S. & Lipsitch, M. Niche and neutral effects of acquired immunity permit coexistence of pneumococcal serotypes. *Science* **335**, 1376–1380 (2012).
- Mammen, M. P. et al. Spatial and temporal clustering of dengue virus transmission in Thai villages. *PLoS Med.* **5**, e205 (2008).
- Salje, H. et al. Revealing the microscale spatial signature of dengue transmission and immunity in an urban population. *Proc. Natl Acad. Sci. USA* **109**, 9535–9538 (2012).

Acknowledgements H.S. and D.A.T.C. acknowledge funding from the National Institutes of Health (R01AI114703-01). S.C. acknowledges financial support from the Investissement d'Avenir program, the Laboratoire d'Excellence Integrative Biology of Emerging Infectious Diseases program (grant ANR-10-LABX-62-IBEID), the Models of Infectious Disease Agent Study of the National Institute of General Medical Sciences, and the AXA Research Fund. The manuscript has been reviewed by the Walter Reed Army Institute of Research. There is no objection to its presentation and/or publication. The opinions or assertions contained herein are the private views of the author, and are not to be construed as official, or as reflecting the true views of the Department of the Army or the Department of Defence. C.K., B.T., A.N., A.W., D.E., I.-K.Y., L.M., R.J., S.T., A.M. and T.E. acknowledge funding from the National Institutes of Health (P01 AI034533) and the Military Infectious Disease Research Program. The investigators have adhered to the policies for protection of human subjects as prescribed in AR 70–25.

Reviewer information Nature thanks P. B. Gilbert, S. Hay and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions H.S., D.A.T.C. and S.C. developed the methods, performed analyses and co-wrote the paper, T.E. conceived the cohort study, T.E., C.K., B.T., A.N., A.W., D.E., L.M., I.-K.Y., R.J., S.T. and A.L.R. ran the study and collected and stored the cohort study results, I.R.-B., J.L. and L.C.K. aided in the interpretation of the results. All authors commented on and edited the paper.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0157-4>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0157-4>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to H.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Cohort study design. Individuals attending 12 different schools in the Kamphaeng Phet district, a rural region of Northern Thailand were recruited into a dengue cohort study that ran between 1998 and 2003 as previously described³². All individuals were between 7 and 13 years old. Blood samples were taken four times a year (in January, June, August and November) with an average of 91 days between blood draws. In addition, from the start of June to mid-November each year, active surveillance was conducted through school-based surveillance. Children who missed school due to febrile illness had additional acute and convalescent blood draws. Dengue infection was confirmed using RT-PCR on the acute sample, with the infecting serotype also recorded or through antibody detection (IgM ELISA values >40 or haemagglutination inhibition test increases of over four times between acute and convalescent blood draws), in which case the infecting serotype was not known. The date of symptom onset, whether or not the child was hospitalized and whether or not they developed DHF was also recorded. Note that the cohort study was conducted before 2009, when the WHO provided new guidance for the characterization of different levels of dengue severity.

Antibody measurements. For each blood draw of each individual, antibody titres for DENV1, DENV2, DENV3 and DENV4 were measured using a haemagglutination inhibition assay. The following twofold dilutions were used: 1:10, 1:20, 1:40, 1:80, 1:160, 1:320, 1:640, 1:1,280 and 1:2,560. We translated each titre onto a log₂ scale such that 1:10 was given a value of 1, 1:20 of value of 2 and so on. Undetectable titres (those with a titre of <1:10) were given a value of 0. For a subset of 800 individuals, 1,771 samples were also tested using PRNTs. These samples were either paired samples from individuals with symptomatic, confirmed infection with one sample taken from a time point before symptom onset and one sample after symptom onset ($n=75$ pairs) or randomly chosen sequential blood samples from individuals without a detected symptomatic infection between the blood draws.

Characterizing how titres change after symptomatic infection. We wanted to understand how titres to both the infecting serotype and to non-infecting serotypes changed over time before and after symptom onset. For all individuals that experienced a symptomatic illness for which the infecting serotype was identified, we identify all titre measurements within each 10-day window from 100 days before symptom onset to 600 days after symptom onset. For each window, we calculated the mean titre of the infecting serotype and the average of the mean titres of the other three serotypes across all individuals that had a blood draw within that window.

Modelling the dynamics of dengue antibody titres. Previous studies on malaria have used hidden Markov models to include undetected infections in estimates of the transmission intensity using presence/absence of specific antibodies in longitudinal data³³. Although these efforts were able to improve estimates of the infection strength within a community compared to using symptomatic individuals, they did not incorporate the changing dynamics of antibody titres over time. By specifically including titre dynamics, we can help to understand a wide range of issues, including assay error, measures of protection and risk and cohort design. **Notation.** We consider an individual i . We denote the number of times the individual was infected before time t . Each dengue infection of individual i is labelled by the index $\psi = 1, \dots, n_i^I(t)$. We denote as $\tau_{i,\psi}^I$ the time of infection number ψ of individual i and $s_{i,\psi}$ is the infecting serotype of infection number ψ of individual i . The history of infection (that is, the timing and serotype of all infections since birth) of individual i up to time t is labelled $H_i(t)$. We denote as N_i^A the total number of times the individual had blood taken during the study. Each blood draw of individual i is labelled by the index $\pi = 1, \dots, N_i^A$. We denote as $\tau_{i,\pi}^A$ the time of blood draw π for individual i . We denote as $A_{i,s,\pi}$ the true antibody titre (see 'Measurement model') and $A_{i,s,\pi}^*$ is the measured antibody titre for individual i for serotype s at blood draw π . $A_i(t)$ represents the cumulative infection strength exerted on individual i before time t . The parameter vector is denoted by θ .

Hierarchical structure of the model. We can break down the probability of a measured antibody titre into three components:

$$P(A_{i,s,\pi=k}^* | A_{i,s,\pi=k}) \\ P(A_{i,s,\pi=k} | \tau_{i,\psi=1}^I, \dots, \tau_{i,\psi=n_i^I(t=\tau_{i,\pi=k}^A)}^I, s_{i,\psi=1}, \dots, s_{i,\psi=n_i^I(t=\tau_{i,\pi=k}^A)}^I, A_{i,s,\pi=1}, \dots, A_{i,s,\pi=k-1}) \\ P(H_i(t=\tau_{i,\pi=k}^A) | \{\lambda\}_t)$$

The first part represents the 'measurement model', the second part the 'antibody dynamics model' and the third part the 'infection model'.

Measurement model. We model the underlying antibody levels on a continuous scale, however, the haemagglutination inhibition assay is a discrete assay, such that in a situation of no measurement error or systematic biases, a true antibody titre between any two dilutions would be measured as the lower of the two dilutions.

So, for example, a true titre of 2.7 would be measured as 2 (assuming there are dilutions performed at 0, 1, 2, 3, ...). In addition, a measurement error is also likely to exist and there may be underlying differences between serotypes (that is, serotype-specific biases) in the assay that will impact all measurements of antibodies against a particular serotype. We consider a 'true titre' to represent the underlying (but unmeasured) titre on a continuous scale. A 'measured titre' is the value that is actually measured by the assay. Conditional on the history of infection of an individual, we assume independence between the measurements of the different serotypes. This seems a reasonable assumption as assays are performed separately for each serotype. The probability of the measured titres ($A_{i,s,\pi=k}^*$) is:

$$P(A_{i,s,k}^* | A_{i,s,k}) = \int_{A_{i,s,k}^*}^{A_{i,s,k}^*+1} f(u) du$$

where $f(u)$ is the density for a normal distribution with mean $A_{i,s,k} + \chi_s$ and a standard deviation parameter (σ), where

$$\chi_s = 0 \text{ if } s = \text{DENV1}$$

$$\chi_s = \chi_2 \text{ if } s = \text{DENV2}$$

$$\chi_s = \chi_3 \text{ if } s = \text{DENV3}$$

$$\chi_s = \chi_4 \text{ if } s = \text{DENV4}$$

Antibody dynamics model. If an individual i was never infected by dengue, we assume that they will have titres of 0 against the four serotypes (this assumes that any maternal antibodies have disappeared and there is no impact of infections by other flaviviruses). At each time point that the individual becomes infected, their antibody titres will increase. We assume that this increase can be broken down into a permanent increase (representing antibodies that will continue to circulate, long after the infection has passed) and a temporary increase (representing the short-lived antibodies generated upon infection).

The permanent rise in titres ($Q_{i,s}(\psi)$), for serotype s from infection number ψ in individual i is modelled as:

$$Q_{i,s}(\psi) = \omega_{i,\psi} K(\psi, s)$$

where $\omega_{i,\psi}$ is a random effect that is gamma-distributed with mean parameter ω_m and variance parameter ω_v and $K(\psi, s)$ allows a differential antibody response for each serotype for primary infections: $K(\psi, s) = \eta$ if it is a primary infection (that is, $\psi = 1$) and s is the infecting serotype; $K(\psi, s) = 1$ otherwise.

We assume that temporary antibody responses will decay exponentially over time:

$$R_{i,s}(t | H_i(t)) = \gamma_{i,\psi=n_i^I(t)} \exp(-(t - \tau_{i,\psi=n_i^I(t)}^I) \delta_{i,\psi=n_i^I(t)}) K(\psi = n_i^I(t), s)$$

where $\gamma_{i,\psi=n_i^I(t)}$ is a random effect that captures the instantaneous rise in temporary antibody titres following the most recent infection (infection $n_i^I(t)$) before time t that comes from a gamma distribution with mean parameter γ_m and variance parameter γ_v ; $\delta_{i,\psi=n_i^I(t)}$ is the rate of decay of the temporary antibodies and comes from a gamma distribution with mean parameter δ_m and variance parameter δ_v . As with the permanent rise in titres, $K(\psi = n_i^I(t), s)$ allows differential antibody responses for primary infections: $K(\psi, s) = \eta$ if it is a primary infection (that is, $\psi = 1$) and s is the infecting serotype; $K(\psi, s) = 1$ otherwise. Additional work is needed to understand whether alternative functional forms for the rise and decay in antibody titres may further refine how antibodies behave following infection.

Under these assumptions and an additional linearity assumption that the temporary and permanent rises are additive, antibody titres at blood draw k for serotype s in individual i is:

$$A_{i,s,\pi=k} = Q_{i,s}(\psi = 1) + \dots \\ + Q_{i,s}(\psi = n_i^I(t = \tau_{i,\pi=k}^A)) \\ + R_{i,s}(t = \tau_{i,\pi=1}^A | H_i(t = \tau_{i,\pi=k}^A)) + \dots \\ + R_{i,s}(t = \tau_{i,\pi=k}^A | H_i(t = \tau_{i,\pi=k}^A))$$

Infection history model. We first assume that both the number of infections and the timing of infections are known. This assumption will subsequently be relaxed. We assume that each individual can get infected up to four times (once by each serotype). The history of infection of an individual depends on seasonality in

dengue transmission and differences in infection intensity across years. For a particular time t , the force of infection is assumed to be:

$$\lambda(t) = \bar{\lambda} \beta_{|t|} \left(1 + \delta \cos \left(\zeta + \frac{2\pi t}{365} \right) \right)$$

where $\bar{\lambda}$ is a parameter that represents the mean daily force of infection in 1998 (the first year of the study) and $\beta_{|t|}$ is the mean force of infection in year $|t|$ compared to the force of infection in 1998.

For an individual i , the contribution to the likelihood for periods before any infection, the probability of their infection history can be broken down into periods of infection and periods without infection. Individuals only contribute to the likelihood during their time in the study.

For each infection that occurs at time t , the contribution to the likelihood is:

$$\log(1 - \exp(-\lambda(t)))$$

For each individual, each day during which no infection occurs, the contribution to the likelihood with respect to serotype s is $\exp(-\lambda(t))$ where more than 90 days have passed since an infection by any serotype and the individual has not previously been infected by serotype s or is 0 otherwise, including periods when the individual is not part of the study

The presence of the 90-day window during which no infection can take place avoids there being more than one infection event between two blood draws. This period is substantially shorter than the estimated period of cross-protection between serotypes of two years³⁴.

In the context of full observation, the probability of the history of infection for individual i can be given as:

$$\begin{aligned} P(H_i(t = \tau_{i,\pi=k}^A) \mid \{\lambda\}) &= \prod_{k=1}^{n_i^I(T_i)} \exp \left(- \int_{\tau_{i,k-1}^I}^{\tau_{i,k}^I} \lambda(u) \, du \right) (1 - \exp(-\lambda(\tau_{i,k}^I))) \\ &\quad \exp \left(- \int_{\tau_{i,n_i^I(T_i)}^I}^{T_i} \lambda(u) \, du \right) \end{aligned}$$

where $\tau_{i,0}^I$ represents the time of birth and T_i the time point at which individual i leaves the study (defined as the day of their final blood draw). We assume the same $\lambda(t)$ for all serotypes.

Situation of imperfect observation. In practice, we do not know the infection history of all individuals. Many infections will have occurred before individuals entered the study. In addition, there are likely to be many subclinical infections that would not have been detected through active surveillance. In addition, active surveillance only operated for 5.5 months of every year. Infections outside these periods would also have been missed (irrespective of symptoms).

For the infection history of individuals before they enter into the study, we estimate a baseline titre ($A_{i,s}(t_0)$) that represents the titre for serotype s one year before the first blood draw. As we assume linearity, such that the temporary and permanent titres of successive historic infections sum up to give the titre at a moment in time, this estimated baseline titre allows us to incorporate the impact of historic infection events up to one year before enrolment but means that we do not need to infer infection events before that time. Individuals that are naive at baseline (defined as those with no measured titres for any serotype at the first blood draw) are given a baseline titre of 0. For an individual with no infection events during the study period, $A_{i,s}(t) = A_{i,s}(t_0)$ for all t .

In the context of full observation during the study period, each individual would have the serotype and time from each infection, $\{s_{i,\psi}, \tau_{i,\psi}\}$, known. For undetected infections or detected infections for which the infecting serotype is unknown (such as when symptomatic infections are only detected through IgM ELISA and therefore the serotype is unknown), we can use a Bayesian data augmentation framework. In this framework, the incompletely observed $\{s_{i,\psi}, \tau_{i,\psi}\}$ pairs are incorporated and considered as nuisance parameters. The joint posterior distribution of the parameters and the augmented data are explored via reversible-jump Markov chain Monte Carlo (MCMC) sampling.

If we call $y = \{s_{i,\psi}^*, \tau_{i,\psi}^*\}_{i=1, \dots, N, \psi=1, \dots, n_i^*(t=0)}$ the observed data,

$z = \{s_{i,\psi}, \tau_{i,\psi}\}_{i=1, \dots, N, \psi=1, \dots, n_i(t=0)}$ the full data (made up of the observed data

and the augmented data), the joint posterior is:

$$P(z, \theta | y) \propto P(y | z) P(z | \theta) P(\theta)$$

$P(y | z)$ represents the observation model, $P(z | \theta)$ is the titre model outlined above and $P(\theta)$ gives the prior distribution of the parameters.

The observation model makes sure that the augmented datasets are consistent with the observed data by having a value of 1 (if consistent) or 0 (if inconsistent). Consistent augmented data have the following characteristics: (1) no individual is infected during the study period by the same serotype more than once; (2) no individual is infected more than once during a 90-day period. Note that, as DENV titre responses to non-DENV flaviviruses, such as Zika and Japanese encephalitis, are likely to be smaller those to DENV infections, such exposures are unlikely to be detected by our model and incorporated as a measurement uncertainty instead.

For all detected (symptomatic) infections, we only detect the date of symptom onset and not the date of infection. To obtain the day of infection for symptomatic cases, we subtract a fixed period of seven days from the day of symptom onset, representing the median incubation period for dengue³⁵. Titres may also not rise on the day of symptom onset (due to recall bias in when symptoms started or individual level variability). For symptomatic infections, we approximate the true, unobserved day of titre increase using augmentation, for which we define consistent augmented data for which the day of titre increase is within ten days of the reported date of symptom onset. For augmented (undetected) infections, we assume that the day of titre rise following infection always occurs 11 days after the day of infection, which represents an approximate estimate of the time between infection and day of titre increase: calculated as the sum of the median incubation period for dengue (seven days) and the median time between symptom onset and titre increase for the detected infections (four days).

This cohort used a rolling recruitment approach, which maintained an approximately constant-sized population and constitutes an important strength compared to cohorts for which the size may be strongly affected by participant dropout. As individuals only contributed to the likelihood for their period of inclusion in the cohort and dropout is not expected to depend on the history of infection, we do not expect that the turnover of participants in the cohort will bias parameter estimates. This was demonstrated in a simulation study in which we were able to recover the true parameters for a simulated cohort with a similar design (see 'Evaluation of the model using simulated data').

We use a log-normal distribution with a log-mean of 0 and log-variance of 1 for the parameters: mean and variance in the permanent rise in \log_2 titres (ω_m, ω_v), mean and variance in the temporary rise in \log_2 titres (γ_m, γ_v), mean and variance in the decay in \log_2 titres per day (δ_m, δ_v), difference in rise for infecting compared to non-infecting serotype (primary infection only) (η), measurement error (σ), DENV2-4 bias (χ_2, χ_3, χ_4), daily infection strength in 1998 per serotype (λ), relative infection strengths versus 1998 for 1997 (β_0) and 1999–2002 ($\beta_2 - \beta_3$) and the two seasonality parameters (δ and ζ).

Estimation using MCMC. We develop a MCMC approach to explore the joint posterior distribution of parameters and the augmented data with the following steps:

(1) Metropolis–Hastings update for the model parameters θ in turn with the updates performed on a logarithmic scale. The step size of the proposals was adjusted to obtain an acceptance probability of 20–30%. As the vast majority of infections are undetected, when updating the six parameters that determine the rise and decay of antibodies (namely $\omega_m, \omega_v, \delta_m, \delta_v, \gamma_m, \gamma_v$), we calculate the likelihood using only the titres from one month before and one year after the symptomatic (and therefore detected) infections. This approach assumes that the rise and fall in titres from all infections come from the same distributions, irrespective of symptom status. More work is needed to understand whether, depending on whether or not an infection leads to symptoms, the titre dynamics following that infection change.

(2) For the symptomatic cases, because the day of titre increase may not fall exactly on the recorded day of symptom onset, we use an independence sampler to update the day of titre increase. At each iteration, the day of the titre increase was updated for 100 randomly chosen symptomatic infections. Candidate values were chosen using a uniform distribution between 10 days before and 10 days after the recorded date of symptom onset.

(3) Independence sampler for the identity of the infecting serotype for the 62 symptomatic infections for which the serotype was not identified. At each iteration, the serotype for each of these infections is updated with equal probability across the four serotypes.

(4) Independence sampler for the identity of the infecting serotype for the undetected infections. At each iteration, the serotype of 500 randomly chosen undetected infections is updated with equal probability across the four serotypes.

(5) Independence sampler for the dates of titre increase for undetected infections. At each iteration, the day of infection is updated for 1,000 randomly chosen undetected infections. For each infection, the proposal is a uniform distribution between one year before entry into the study and the day of the final blood draw.

(6) Independence sampler for the baseline titres for each individual. At each iteration, the baseline titre for one serotype is updated for 1,000 randomly chosen

individuals. The proposal distribution is a random uniform distribution between 0 and 10. All individuals that are naive at baseline (that is, those with no titres to any serotype at the first blood draw) are forced to have a baseline titre to 0 for all four serotypes.

(7) Reversible-jump MCMC to add/remove unobserved infection events. As $H_i(t_i)$ is unobserved, we use a Bayesian data augmentation approach that treats it as a nuisance parameter. Rather than attempting to definitively identify whether an infection occurred or not, these approaches allow us to incorporate the uncertainty of the presence and timing of these events. We use a reversible-jump MCMC model to add and remove infection events. Each step to add undetected infections proceeds as follows:

- (i) Randomly draw individual.
 - (ii) Draw a candidate date for the infection event using a uniform distribution from one year before their first blood draw to the day of their final blood draw.
 - (iii) Draw a candidate serotype of infection with the probability of each serotype being 0.25.
 - (iv) Update the number, date and serotype of infections for that individual.
- For the removal of undetected infections, we use a similar approach:
- (i) Randomly draw individual.
 - (ii) If that individual has undetected infections, randomly select one of their infections with equal probability (if they have no infections move to the next individual).
 - (iii) Update the number, date and serotype of infections for that individual by removing that infection.

Evaluation of the model using simulated data. In order to evaluate the ability of the model to accurately estimate the parameters in a scenario when only a minority of infections are observed, we use the same modelling framework on a random subset of 1,000 individuals from the study with subsequent changes in titres. We include the actual start date and the end date for these individuals (that is, when they entered and left the cohort). We simulate infections in these individuals based on known parameters. We then randomly ‘unobserve’ 70% of infections to reflect undetected infections. We then estimate the parameters using our framework and compare them to the underlying true parameters.

Sensitivity analysis using school-specific infection strength parameters. The infection strength exerted on individuals may differ across schools, resulting in non-independence between individuals attending the same school. To assess the impact of any such correlation on our parameters, we performed a sensitivity analysis in which we included a separate force of infection parameter for each school. In this model the force of infection exerted on an individual that attends school (sch) is:

$$\lambda(t, \text{sch}) = \bar{\lambda} \beta_{|t|} \beta_{|\text{sch}|} \left(1 + \delta \cos \left(\zeta + \frac{2\pi t}{365} \right) \right)$$

where $\bar{\lambda}$ is a parameter that represents the mean daily force of infection in 1998 in school 1, $\beta_{|t|}$ is the mean force of infection in year $|t|$ as compared to that in 1998 and $\beta_{|\text{sch}|}$ is the mean force of infection for school sch compared to school 1.

Alternative functional forms for the decay in titres. Alternative functional forms for the decay in antibody titres exist. In particular, biphasic models that model both short-term and longer-term antibody decay with different exponential decay rates have been shown to work well in other systems, such as malaria³⁶. The biphasic form is captured by:

$$\text{Titre}_i = \theta_1 \theta_2 \exp(-\theta_3 t) + (1 - \theta_2) \exp(-\theta_4 t)$$

where θ_1 , θ_2 , θ_3 and θ_4 capture the decay of the titres. To explore whether this biphasic form may further refine how antibodies behave following infection, we fitted both exponential decay and biphasic models to the observed infections using the observed titres following detected PCR-confirmed infections and the dates of symptom onset. We found largely consistent results in the two models (Extended Data Fig. 1). As exponential decay is the more parsimonious model, we retained this form for the final analysis. Nevertheless, structural uncertainty in the model used for the analysis remains, which will not be represented within the confidence intervals for the parameters.

Estimation of impact on titres on infection and disease. We use the augmented times and serotypes of infection from 100 model iterations to reconstruct the antibody titre trajectories for each individual. For each augmented dataset, we extract the mean titre across all four serotypes for each day and whether they got infected the following day or not. Person time for individuals who were considered not susceptible (that is, had been infected in the prior 90 days) was excluded. To explore the relationship between mean titre and the probability of infection, we conducted logistic regression for which a polynomial spline of order 2 for the mean titre was used (determined as the optimal model through comparison of different polynomial models using the Akaike information criterion (AIC)). To account for sampling uncertainty, in each reconstructed dataset, we use a bootstrap approach to sample all individuals with replacement and then re-perform the logistic regression

each time. We present the mean and 95% confidence intervals from the resulting distribution of the logistic model estimates of the probability of infection for each titre obtained from across the model iterations.

We explore the relationship between mean titre and the probability of having different disease outcomes. We consider three different outcomes: symptomatic infection (irrespective of severity), hospitalization and DHF. We only consider titres during the active surveillance windows and whether or not individuals had an infection the following day that led to the outcome of interest. For each outcome, we conduct logistic regression in which we use a polynomial spline of order 2 for the mean titre (consistently determined as the optimal model through comparison of different polynomial models using AIC). We use a bootstrap approach to sample all individuals with replacement and then re-perform the logistic regression each time and identified the mean and 95% confidence intervals from the resultant distribution for the estimates of the probability of having an infection that led to the outcome of interest for each titre obtained from across the model iterations.

For those individuals that became infected during the active surveillance windows, we fit logistic models to the mean titres and whether or not the disease outcome occurred. We looked at three outcomes: any symptomatic illness, hospitalization and DHF. For each of the three outcomes, we compare an intercept-only model with models with a polynomial spline up to order 2. To account for sampling uncertainty, in each reconstructed dataset, we use a bootstrap approach to sample all individuals who had an infection during the surveillance windows with replacement and then re-perform the logistic regression each time. We present the mean and 95% confidence intervals from the resultant distribution of the logistic model estimates of the probability of infection for each titre obtained from across the model iterations.

PRNT titres are available for a subset of 1,771 blood draws. For those that became infected during the active surveillance windows and PRNT titres are available in the six-month window before infection, we fit logistic models to these mean PRNT titres from that six-month time frame and whether or not the disease outcome occurred. We looked at three outcomes: any symptomatic illness, hospitalization and DHF. For each of the three outcomes of interest, we compare an intercept only model with models with a polynomial spline up to order 2. To account for sampling uncertainty, in each reconstructed dataset, we use a bootstrap approach to sample all individuals who had an infection during the surveillance windows with replacement and then re-perform the logistic regression each time. To account for the fact that individuals and serum samples may not have been completely selected at random for PRNT testing (for example, preferential testing of those with symptomatic disease), we adjusted our estimate for the probability of sampling conditional on the outcome of interest.

From the logistic regression described above, we can extract the probability of the outcome of interest given a particular PRNT titre and that a PRNT was conducted. Using Bayes rule we can write down:

$$P(\text{outcome} | \text{titre, PRNT done}) = \frac{P(\text{PRNT done} | \text{outcome, titre}) P(\text{outcome} | \text{titre})}{P(\text{PRNT done} | \text{titre})}$$

as the PRNT titre (or the haemagglutination inhibition titre) was not taken into account in the selection process for choosing whether or not a PRNT was done, this becomes:

$$P(\text{outcome} | \text{titre, PRNT done}) = \frac{P(\text{PRNT done} | \text{outcome}) P(\text{outcome} | \text{titre})}{P(\text{PRNT done})}$$

As we are interested in $P(\text{PRNT done} | \text{outcome})$, we can reorder this equation to:

$$P(\text{PRNT done} | \text{outcome}) = \frac{P(\text{outcome} | \text{titre, PRNT done}) P(\text{PRNT done})}{P(\text{outcome} | \text{titre})}$$

We therefore multiply our logistic model outcomes by the following adjustment factor:

$$\text{adjustment factor} = \frac{P(\text{PRNT done})}{P(\text{PRNT done} | \text{outcome})}$$

$P(\text{PRNT done})$ is calculated as the proportion of all infection events for which a PRNT was conducted in the prior 6 months from the infection and $P(\text{PRNT done} | \text{outcome})$ is calculated as the proportion with the outcome of interest for which PRNTs were conducted in the prior 6 months. We present the mean and 95% confidence intervals from the resultant distribution of the logistic model estimates of the probability of infection for each titre obtained from across the model iterations.

We used a logistic regression approach to explore the impact of year of infection and the age at the time of infection. To explore the impact of year, we take each augmented dataset in turn and sample all the individuals with replacement to incorporate sampling uncertainty. We then regress the year of infection (as a categorical variable) on whether the outcome $Y_{i,t}$ occurred:

$$\text{logit}(Y_{i,t}) = \beta_0 + \beta_1 \times \text{Year}_{i,t}$$

where $\text{Year}_{i,t}$ is the year (1998, 1999, 2000, 2001 or 2002) within which day t occurred for individual i . We conducted separate regressions for which the outcome was an infection event (irrespective of whether the infection led to symptoms), symptomatic infection events (irrespective of disease severity), hospitalization and development of DHF. For the last three models, we only considered data during the active surveillance windows, as we do not know the symptom status of infections outside these windows. To explore the impact of age, we dichotomized the age of individuals as being less than or greater than 9 (the Sanofi–Pasteur vaccine is not recommended for individuals under 9). We then performed the regression:

$$\text{logit}(Y_{i,t}) = \beta_0 + \beta_1 \times \text{Age}_{i,t}$$

where separate models for the same four outcomes, $Y_{i,t}$, were performed. Finally, we built multivariable models that also accounted for mean titre using a polynomial of order 2:

$$\text{logit}(Y_{i,t}) = \beta_0 + \beta_1 \times \text{Age}_{i,t} + \beta_2 \times \text{Titre}_{i,t} + \beta_3 \times \text{Titre}_{i,t}^2$$

Impact of titre on outcome using Cox proportional hazard models. In the context of small probabilities of an event occurring and short time intervals between readings, logistic regression will give consistent results with those obtained from Cox proportional hazards models that specifically takes the non-independence of titre observations from the same individuals into account¹⁷. To demonstrate the consistency of the two approaches we estimate the impact of titre on our four outcomes (infection, symptomatic infection, hospitalized infection and DHF infection) using a time-varying Cox proportional hazards model, specifically incorporating clustering of observations by individual³⁷. We used 100 augmented datasets. For each augmented dataset, we extract the mean titre across all four serotypes for each day and whether they got the outcome of interest the following day or not. For the disease-specific outcomes (any symptomatic disease, hospitalized infection and DHF infection), we only used time points during the surveillance windows. We then calculated the impact of the mean titre (polynomial of order 2) on the relative hazard of infection, incorporating a clustering ID per individual using the survival package in R³⁷. We then calculate the mean effect of titre on the outcome of interest by averaging the estimates across the reconstructed datasets.

To compare our results using logistic regression, we multiply the annualized estimate of a titre x on the risk of the outcome (calculated as $1 - \exp(-365x)$) by the estimated baseline hazard for those cases with a measured titre of 0 (calculated as the proportion of infections in time points with a measured titre of 0). We find that the results are almost identical (Extended Data Fig. 6). As the logistic model approaches allow us to directly estimate the underlying probability of the outcome, it is preferred.

Survival analysis. Annualized probability of infection using titre data only. Over 100 reconstructed datasets, we initially identify all individuals who experienced an infection (irrespective of disease severity). We then identify the set-point antibody load for that infection as the mean titre one year following infection as predicted by our model. Individuals were divided into two groups, those with a set-point antibody load ≤ 3 and those with a load > 3 . For each individual in each titre group, we use the logistic model described in ‘Estimation of impact on titres on infection and disease’ to predict the daily probability of a subsequent infection based on the mean titres each day following the initial infection. We also calculated the daily probability of experiencing an infection that leads to DHF. We annualize the predicted probabilities of subsequent infection by using the conversion $1 - \exp(-365x)$ where x is the daily probability of infection. We present the mean annualized probabilities across all individuals and over all the reconstructed datasets.

Kaplan–Meier analysis. For individuals who experienced an infection, we calculate Kaplan–Meier survival curves for experiencing a subsequent infection (both irrespective of disease outcome and for DHF only). Over 100 reconstructed datasets, we identify all individuals who experienced an infection event. We then identify the set-point antibody load for that infection as the mean titre one year following infection as predicted by our model. Individuals were divided into two groups, those with a set-point antibody load ≤ 3 and those with a load > 3 . To incorporate sampling uncertainty, we resample all individuals with replacement. For each group, we then calculate Kaplan–Meier survival curves. We present the mean and 2.5 and 97.5 quantiles from the resulting distribution.

Prediction of DHF outcome using mean titre. We assess the ability of our logistic model to discriminate between those who developed DHF and those who did not using a leave-one-out cross-validation method.

For each reconstructed dataset, taking each DHF case in turn, we initially identified all individuals who were in the cohort at the same time as the DHF infection with detectable titres who themselves did not have a DHF infection within a 1-year period. We then randomly selected one of those individuals and used the titre from that day. Once we had selected a matched control for each DHF case, we calculated the receiver operating characteristic (ROC) using leave-one-out cross-validation. To do this, we removed each individual in turn from the dataset (including both the cases and the controls) and recalculated the relationship between mean haemagglutination inhibition titre and DHF infection using all the remaining titre readings. We then predicted the probability that the held-out case had a DHF infection. The ROC was calculated using these probabilities across individuals. We present the mean ROC from across 100 reconstructed datasets.

Clustering of infections by school. For additional model validation, we explore whether augmented infections occurred in the same schools around the same time as observed cases, despite no information on location being provided to the model.

Clustering of subclinical infections within schools. To explore the clustering of subclinical data with symptomatic infections in schools, we use the tau-clustering statistic^{31,38} to calculate the odds of observing an subclinical infection (irrespective of serotype and infection parity) within a set time period (t_1, t_2) of a symptomatic infection within the same school relative to the odds of observing a subclinical infection in a different school within the same time window.

$$\hat{\tau}(t_1, t_2) = \frac{\hat{\pi}(t_1, t_2)}{\hat{\pi}(\infty)}$$

where:

$$\hat{\pi}(t_1, t_2) = \frac{\sum_{i=1}^{N_{\text{symp}}} \sum_{j=1}^{N_{\text{asymp}}} \mathbf{I}(\text{sch}_{ij} = 1, t_1 < |s_{ij}| < t_2)}{\sum_{i=1}^{N_{\text{symp}}} \sum_{j=1}^{N_{\text{asymp}}} \mathbf{I}(\text{sch}_{ij} = 0, t_1 < |s_{ij}| < t_2)}$$

where N_{symp} and N_{asymp} are the number of symptomatic and subclinical infections within any model iteration, \mathbf{I} is an indicator variable, sch_{ij} is equal to 1 if individuals i and j go to the same school and 0 otherwise, s_{ij} is the time between infections. We varied the time window between 0–90, 90–180 days and greater than 180 days.

Clustering of serotypes within schools. We explore whether the augmented serotypes that were assigned to subclinical primary infections (serotypes could not reliably be assigned in subsequent infections because of cross-reactions) were consistent with the serotypes of the symptomatic infections of individuals within the same school for different periods of time.

For augmented primary infections that are consistently of the same serotype (defined as $> 50\%$ of augmented datasets having a primary infection in the same individual caused by the same serotype in the same six-month time window), we calculated the odds that an augmented primary infection that occurs in the same school and within a fixed time window of a PCR-confirmed case is of the same serotype relative to the odds that an augmented primary infection that occurs within the same time window in a different school is of the same serotype.

$$\hat{\tau}_2(t_1, t_2) = \frac{\hat{\pi}_2(t_1, t_2)}{\hat{\pi}_3(t_1, t_2)}$$

where:

$$\hat{\pi}_2(t_1, t_2) = \frac{\sum_{i=1}^{N_{\text{symp}}} \sum_{j=1}^{N_{\text{asymp}}} \mathbf{I}(\text{sch}_{ij} = 1, t_1 < |s_{ij}| < t_2, \text{ser}_{ij} = 1)}{\sum_{i=1}^{N_{\text{symp}}} \sum_{j=1}^{N_{\text{asymp}}} \mathbf{I}(\text{sch}_{ij} = 1, t_1 < |s_{ij}| < t_2, \text{ser}_{ij} = 0)}$$

$$\hat{\pi}_3(t_1, t_2) = \frac{\sum_{i=1}^{N_{\text{symp}}} \sum_{j=1}^{N_{\text{asymp}}} \mathbf{I}(\text{sch}_{ij} = 0, t_1 < |s_{ij}| < t_2, \text{ser}_{ij} = 1)}{\sum_{i=1}^{N_{\text{symp}}} \sum_{j=1}^{N_{\text{asymp}}} \mathbf{I}(\text{sch}_{ij} = 0, t_1 < |s_{ij}| < t_2, \text{ser}_{ij} = 0)}$$

where ser_{ij} is equal to 1 if i and j go to the same school and 0 otherwise. We varied the time window between 0–90 days, 90–180 days and greater than 180 days.

Uncertainty. To incorporate sampling uncertainty into our estimates, for each model iteration, we randomly selected all infection events with replacement before calculating the tau estimates. The 95% confidence intervals were calculated from the 2.5% and the 97.5% quantiles of the resulting distribution across all model iterations.

Different approaches to identify infections using simple cutoff points. To assess the sensitivity and specificity of the current approach to identify infections based on titre differences across two blood draws, we simulated titre trajectories in which infections did and did not take place.

Simulated titres in which infections did take place. We used the following algorithm:

- (1) Randomly draw MCMC iteration.
- (2) Randomly divide the population of individuals who had at least one infection in two: 'model fit' individuals and 'held out' individuals.
- (3) Of the model fit individuals, randomly draw an individual i .
- (4) Identify the parameters for the antibody dynamics for the first infection for that individual (that is, $\psi_{i,\tau=1}$, $\gamma_{i,\tau=1}$, $\omega_{i,\tau=1}$) and the baseline titre $A_{i,s}(t_0)$ from that MCMC iteration. The true titre for each serotype will be $A_{i,s}(t_0)$.
- (5) Calculate the measured titre for each serotype using a random draw from a normal distribution with mean $A_{i,s}(t_0)$ and standard deviation σ , where σ represents the measurement error of the assay. Under scenarios of a discrete assay, the measured titre is rounded down to the nearest integer.
- (6) Draw an infection time point using a uniform distribution between 0 and t_{\max} , where t_{\max} represents the time of the second blood draw.
- (7) Calculate the true titre at t_{\max} for each serotype ($A_{i,s}(t_{\max})$).
- (8) Calculate the measured titre using a random draw from a normal distribution with mean $A_{i,s}(t_{\max})$ and standard deviation σ . Under scenarios of a discrete assay, the measured titre is rounded down to the nearest integer.

Simulated titres where infections did not take place. We used the following algorithm:

- (1) Randomly draw MCMC iteration.
- (2) Randomly divide the population of individuals who had at least one infection in two: 'model fit' individuals and 'held out' individuals.
- (3) Of the model fit individuals, randomly draw an individual i .
- (4) Identify the baseline titre $A_{i,s}(t_0)$ from that MCMC iteration. The true titre for each serotype will be $A_{i,s}(t_0)$.
- (5) Calculate the measured titre for each serotype using a random draw from a normal distribution with mean $A_{i,s}(t_0)$ and standard deviation σ , where σ represents the measurement error of the assay. Under scenarios of a discrete assay, the measured titre is rounded down to the nearest integer.
- (6) Calculate a second measured titre using a random draw from a normal distribution with mean $A_{i,s}(t_0)$ and standard deviation σ . Under scenarios of a discrete assay, the measured titre is rounded down to the nearest integer.

Different assays. The current approach is to see whether there is a fourfold rise between blood draws in any of the four serotypes using the discrete haemagglutination inhibition assay.

The 'mean' approach is to first calculate the mean across the four serotypes at each time point and then compare the mean titres across two time points to identify whether infections have occurred or not.

Some assays give titres on a continuous scale (and not discretized like the haemagglutination inhibition assay). In the 'continuous assay' approach, as with the mean approach, we initially calculate the mean titre across the four serotypes at each time point and then compare the mean titres across two time points to identify whether infections have occurred or not.

Assessment of the different assays across time between blood draws and error in assay. Using the simulation approaches set out above we obtained 10,000 individuals with pairs of measured titres (with one titre for each serotype) for whom an infection did take place in between the titre measurements and a further 10,000 individuals with pairs of measurements for whom no infection took place. We varied the time between blood draws (t_{\max}) between 10 days and 400 days and the error in the assay (σ) between 0.1 and 1. For each resultant dataset, we used the held-out dataset (that is, those individuals not included in the model fitting) to calculate the sensitivity and specificity under each of the different measurement approaches. Each time, we also identified the cutoff point that maximized the sensitivity while maintaining at

least 95% specificity. We performed a separate analysis in which we identify cutoff points to maximize sensitivity while maintaining 99% specificity.

Comparison between PRNT and haemagglutination inhibition titres. For 1,771 blood draws, both PRNTs and haemagglutination inhibition tests were conducted. We compared the mean PRNT log₂ titre across the four serotypes with the mean haemagglutination inhibition log titre from the four serotypes and fitted a line through the two using linear regression. We compared different polynomial models up to order 2 and used the best fitting one as determined by AIC.

Comparison with Sanofi–Pasteur vaccine titres. To explore the potential impact of the Sanofi vaccine, we extracted the geometric mean PRNT titres following vaccination for both seronegative and seropositive individuals who were vaccinated in Latin America²¹. The extracted values for PRNT titre, 28 days after the second injection (see Supplementary Table 8 in Villar et al.²¹) are shown in Supplementary Table 8.

The values 28 days after the third injection are also available and are 81 for those who were seronegative before vaccination and 658 for those who were seropositive before vaccination²¹. We plot these values on a plot of the relationship between haemagglutination inhibition titre and PRNT titre from our assays (Fig. 4d).

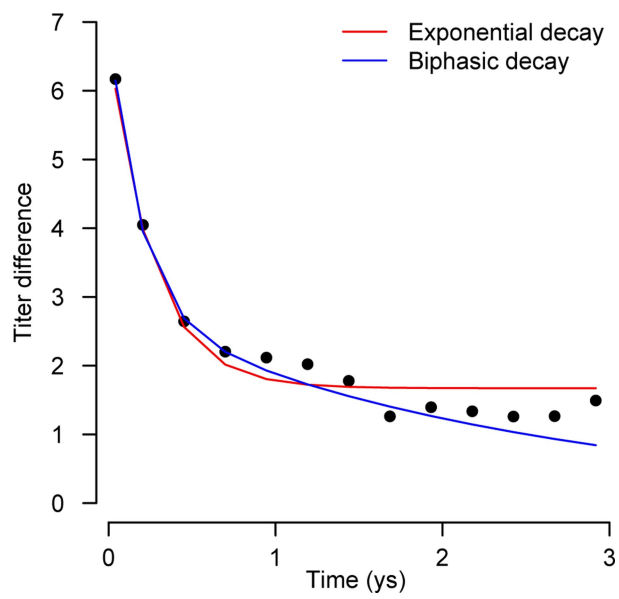
Ethical approval. The cohort protocol was approved by the Institutional Review Boards of the Thai Ministry of Public Health, the Office of the US Army Surgeon General and the University of Massachusetts Medical School. Informed consent was obtained from participants and their parents/guardians. No personal identifiable information was available to the researchers for the presented analysis.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

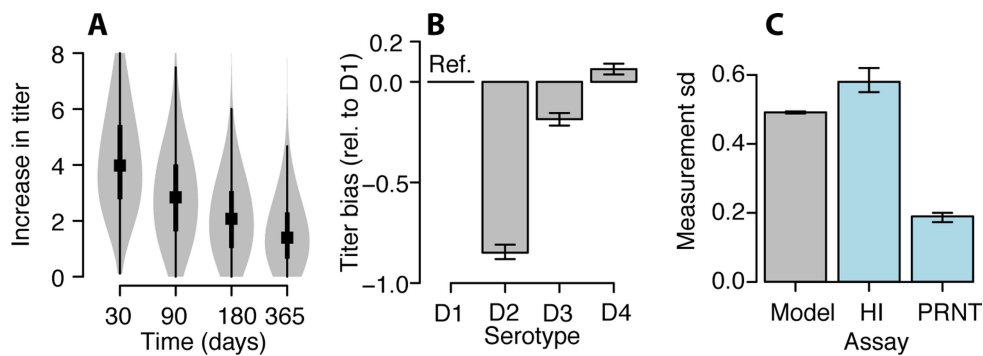
Code availability. C++ code is available from the corresponding author upon request.

Data availability. De-identified data used in this project are available as part of this manuscript. All date information was removed in order to de-identify the dataset. Individuals interested in accessing a full dataset with identifying information should contact the corresponding author to obtain the necessary IRB approval.

32. Endy, T. P. et al. Spatial and temporal circulation of dengue virus serotypes: a prospective study of primary school children in Kamphaeng Phet, Thailand. *Am. J. Epidemiol.* **156**, 52–59 (2002).
33. Smith, T. & Vounatsou, P. Estimation of infection and recovery rates for highly polymorphic parasites when detectability is imperfect, using hidden Markov models. *Stat. Med.* **22**, 1709–1724 (2003).
34. Reich, N. G. et al. Interactions between serotypes of dengue highlight epidemiological impact of cross-immunity. *J. R. Soc. Interface* **10**, 20130414 (2013).
35. Rudolph, K. E., Lessler, J., Moloney, R. M., Kmush, B. & Cummings, D. A. T. Incubation periods of mosquito-borne viral infections: a systematic review. *Am. J. Trop. Med. Hyg.* **90**, 882–891 (2014).
36. White, M. T. et al. A combined analysis of immunogenicity, antibody kinetics and vaccine efficacy from phase 2 trials of the RTS,S malaria vaccine. *BMC Med.* **12**, 117 (2014).
37. Therneau, T., Crowson, C. & Atkinson, E. Using time dependent covariates and time dependent coefficients in the Cox model. R package Survival Vignettes version 2.41 <https://cran.r-project.org/web/packages/survival/vignettes/> (2017).
38. Lessler, J., Salje, H., Grabowski, M. K. & Cummings, D. A. T. Measuring Spatial Dependence for Infectious Disease Epidemiology. *PLoS ONE* **11**, e0155249 (2016).

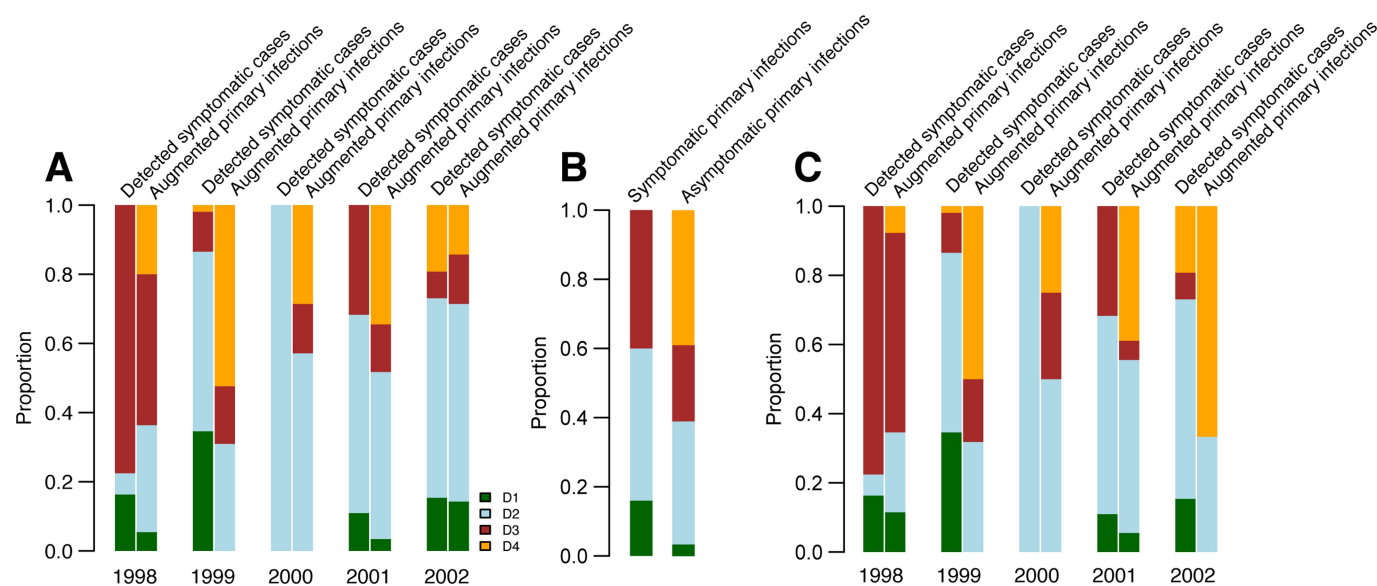


Extended Data Fig. 1 | Comparison of biphasic versus exponential decay. Biphasic and exponential decay curves fitted to haemagglutination inhibition antibody measurements following observed symptomatic infections.



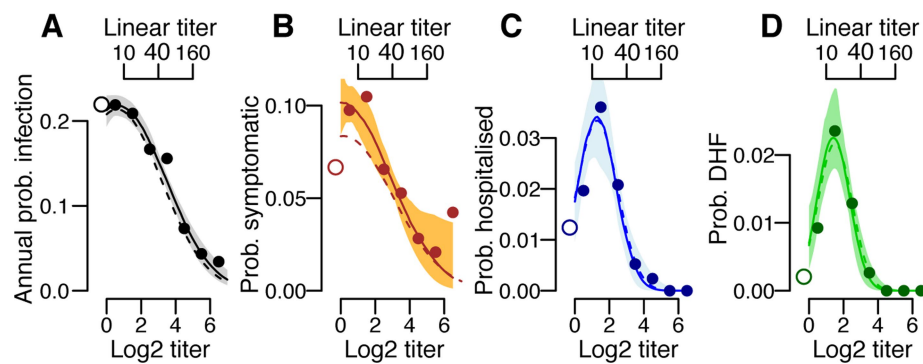
Extended Data Fig. 2 | Variability in titre responses and measurement error and bias by serotype. a, Variability in titre responses. Violin plots showing median (black square), 25% and 75% quantiles (thick black line) and 95% distribution (in grey) of net titre increases at different time points after infection ($n = 1,420$). **b,** Estimated underlying differences across serotypes in the measurement of antibody levels by haemagglutination inhibition assay over that attributable to infection (DENV1 is reference

(Ref.)) with 95% credible intervals (fitted to data from 140,612 titre measurements). **c,** Mean estimated error in the haemagglutination inhibition assay estimated with 95% credible intervals using our model results (grey) and empirically derived (blue) results from 795 repeated measurements on the same serum compared to the values from previously empirically derived estimates¹² for PRNTs (blue).



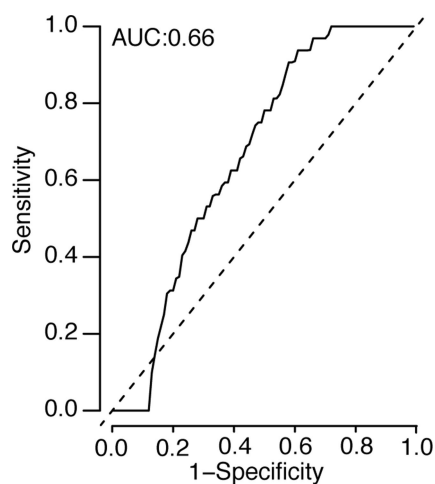
Extended Data Fig. 3 | Serotype distributions. **a**, Distribution of serotypes by year comparing the detected symptomatic infections by PCK and the augmented primary infections for which we could confidently assign the serotype (>50% of model iterations inferring the same serotype). We could confidently assign the serotype in 60% of cases. **b**, Serotype distribution for detected symptomatic primary infections and augmented subclinical primary infections for which the infecting serotype

could be confidently assigned (>50% of model iterations inferring the same serotype). **c**, Distribution of serotypes by year comparing the detected symptomatic infections by PCR and the augmented primary infections using a more stringent cutoff that >75% of model iterations infer the same serotype. In this scenario, we could confidently assign the serotype in 32% of instances.



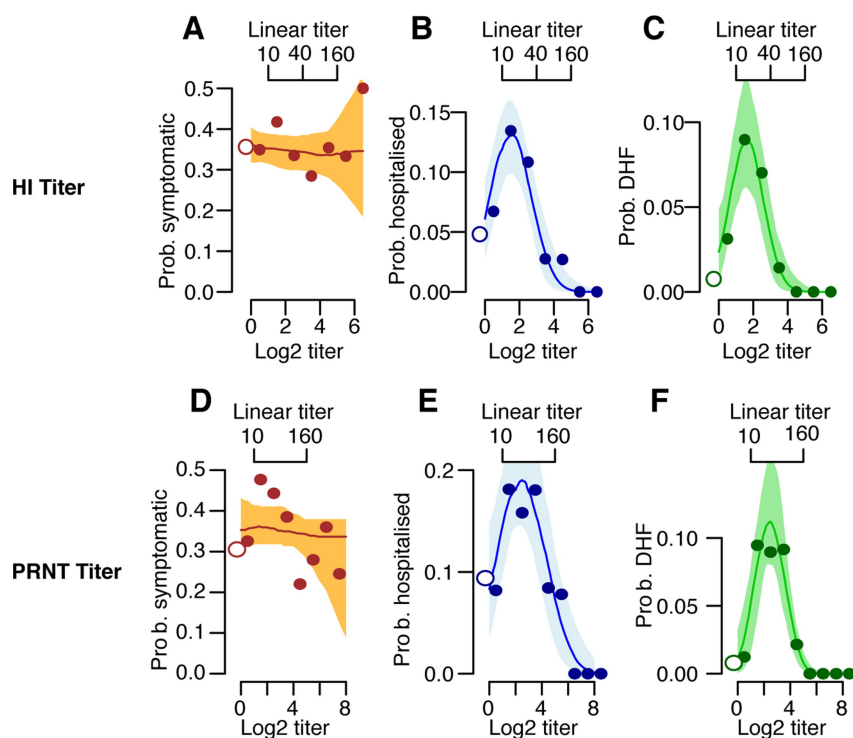
Extended Data Fig. 4 | Cox proportional hazards model versus logistic regression. Comparison of results using time varying Cox proportional hazards model (dashed line) with that from logistic regression (solid line) for the annualized probability of infection (a), developing any symptoms (b), being hospitalized (c) and developing DHF (d) as a function of the mean measured antibody titre across all serotypes at the time of exposure using titre data from all study subjects ($n = 3,451$). The open circles on

the left represent primary infections (that is, those with no detectable titres for any serotype before exposure). The shaded regions represent 95% bootstrap confidence intervals. To calculate probabilities, the relative hazards from the Cox model are multiplied by the baseline hazard for those with measured titres of 0 (calculated as proportion of person time with an infection time among those with measured titres of 0).



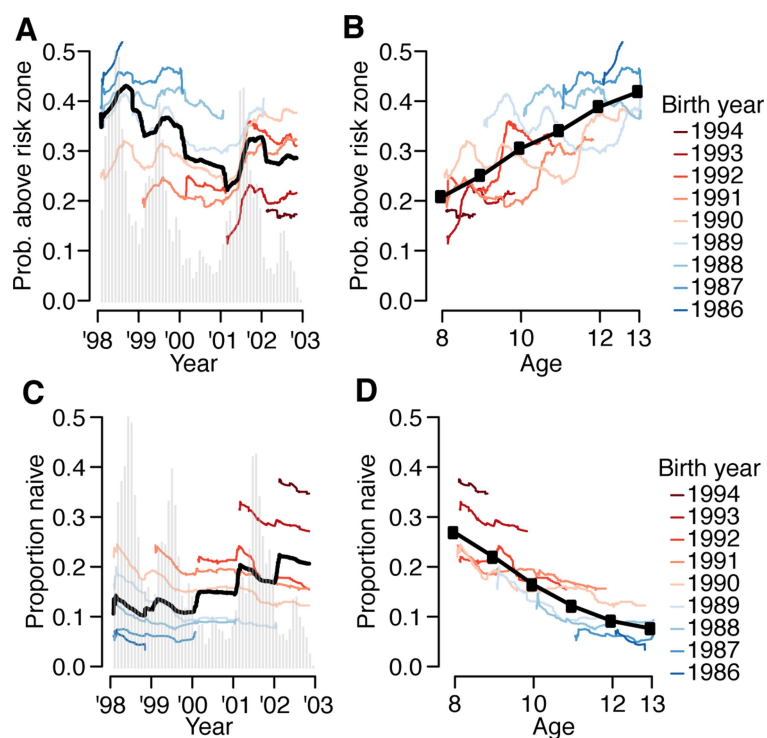
Extended Data Fig. 5 | ROC for the identification of DHF infections.

Ability of modelled relationship between measured haemagglutination inhibition titre and risk of DHF to identify those with DHF, using those with DHF compared to randomly selected matched controls from individuals in the cohort who had detectable titres at the same time ($n = 36$ with DHF with the same number of matched controls). AUC, area under the curve.



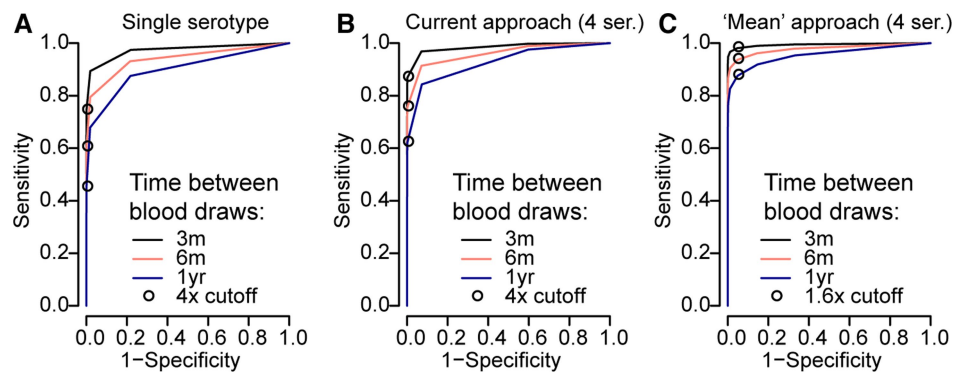
Extended Data Fig. 6 | Probability of disease as a function of haemagglutination inhibition and PRNT titre. Probability of disease as a function of mean titre across the four serotypes at the time of infection. **a**, For those infected during the surveillance windows, the probability of developing any symptoms as a function of mean titre ($n = 781$). **b**, For those infected during the surveillance windows, the probability of being hospitalized ($n = 781$). **c**, For those infected during the surveillance windows, the probability of developing DHF as a function of mean titre

($n = 781$). **d**, For those infected during the surveillance windows ($n = 781$), the probability of developing any symptoms as a function of mean PRNT titre. **e**, For those infected, the probability of being hospitalized as a function of mean PRNT titre. **f**, For those infected, the probability of developing DHF as a function of mean PRNT titre. In each panel, the open circles on the left represent primary infections. The shaded region represents 95% confidence intervals.



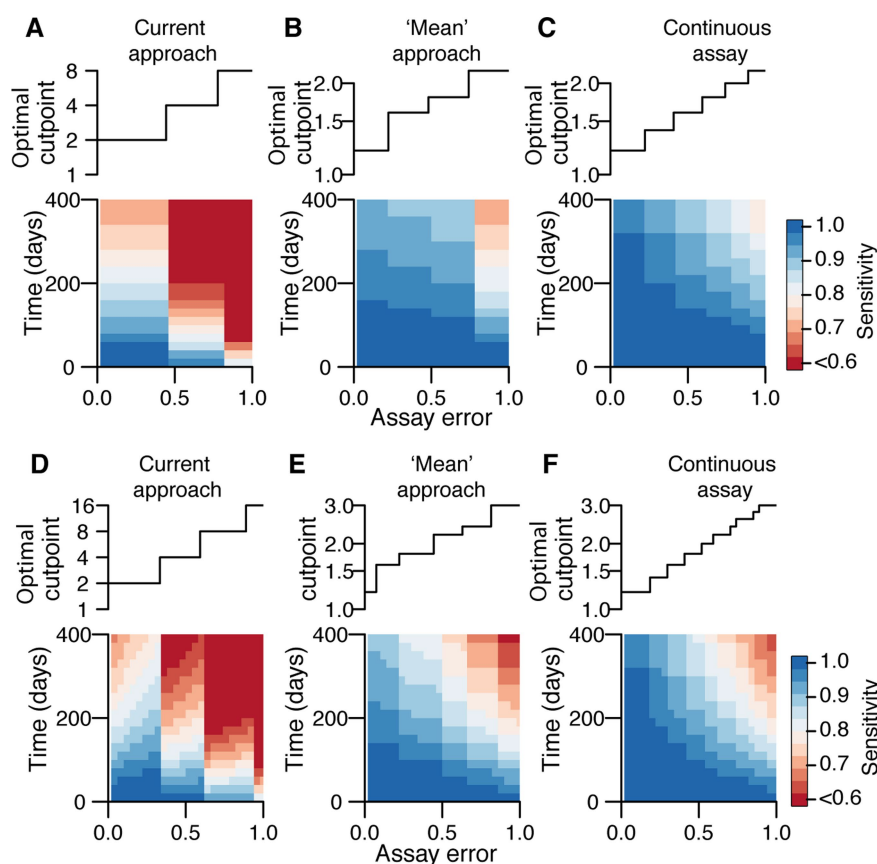
Extended Data Fig. 7 | Population-level distribution of titres by birth cohort and age. **a**, Proportion of each cohort who are naive as a function of time. **b**, Proportion of each cohort who are naive as a function of age.

c, d, Proportion of each cohort with titres above risk zone (that is, greater than 3) as a function of time (**c**) and age (**d**).



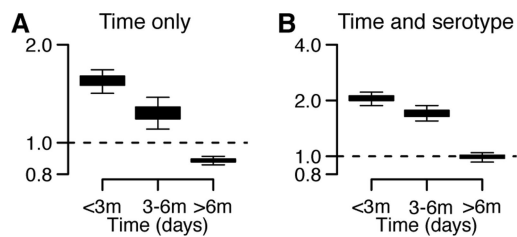
Extended Data Fig. 8 | ROC for infection detection using different testing protocols. The ROC for different assay approaches and time between blood draws calculated from 100,000 simulated titre responses. **a**, Single serotype assay—when haemagglutination inhibition tests are conducted for only a single serotype at two time points. **b**, Haemagglutination inhibition tests conducted against all four serotypes.

Infections are considered to occur when the ratio of any of the four titres at time point 2 versus time point 1 is greater than the threshold value. **c**, Haemagglutination inhibition tests conducted against all four serotypes. Infections are considered to occur when the ratio of the mean of the four titres at time point 2 versus the mean at time point 1 is greater than the threshold value.



Extended Data Fig. 9 | Performance of assay is dependent on time between blood draws and measurement error. Optimization of assays for the detection of events in which the specificity is maintained at >95%. **a–c**, We explore the performance of three different assay testing protocols: current practice for which infection events are defined as a rise above a cutoff point in any serotype across two blood draws (**a**), a ‘mean approach’ for which the mean across all serotypes is first calculated before comparing the means across time points (**b**), a ‘mean approach’ for which titres are

available on a continuous scale (**c**). For each protocol, we identify the optimal cutoff point for a range of assay measurement errors from 100,000 simulated titres based on the fitted titre responses from infections in our study population, that maintains a specificity of >95% (top row). We then calculate the sensitivity of the approach for different time intervals between blood draws using 50% held-out data (bottom row). **d–f**, Same as **a–c** but using a more stringent cutoff of 99%.



Extended Data Fig. 10 | Clustering of symptomatic ($n = 274$) and subclinical cases (mean $n = 507$ across 100 reconstructed datasets) by school by time and serotype. **a, Probability of observing an augmented subclinical infection (irrespective of serotype) occurs at different time intervals within the same school of a detected symptomatic case relative to the probability of observing an augmented subclinical infection occurring in a different school in that same time interval. **b**, For augmented primary infections that are consistently of the same serotype (defined as $>50\%$ of augmented datasets having a primary infection in the same individual caused by the same serotype in the same six-month time window). Probability that an augmented primary infection that occurs within a fixed time window of a PCR-confirmed case and in the same school is of the same serotype relative to the probability that an augmented primary infection that occurs within the same time window in a different school is of the same serotype. Note that the modelling framework can only allow differentiation of serotypes for primary infections. Cross-reaction prevents differentiation in subsequent infections. Overall, 60% of primary infections have a consistent serotype for a primary infection across augmented datasets. Each box plot presents the 2.5%, 25%, 75% and the 97.5% quantiles of the distribution as well as the mean.**

Microglial control of astrocytes in response to microbial metabolites

Veit Rothhammer¹, Davis M. Borucki¹, Emily C. Tjon¹, Maisa C. Takenaka¹, Chun-Cheih Chao¹, Alberto Ardura-Fabregat², Kalil Alves de Lima¹, Cristina Gutiérrez-Vázquez¹, Patrick Hewson¹, Ori Staszewski², Manon Blain³, Luke Healy³, Tradite Neziraj¹, Matilde Borio¹, Michael Wheeler¹, Loic Lionel Dragin⁴, David A. Laplaud⁵, Jack Antel³, Jorge Ivan Alvarez⁴, Marco Prinz^{2,6} & Francisco J. Quintana^{1,7*}

Microglia and astrocytes modulate inflammation and neurodegeneration in the central nervous system (CNS)^{1–3}. Microglia modulate pro-inflammatory and neurotoxic activities in astrocytes, but the mechanisms involved are not completely understood^{4,5}. Here we report that TGF α and VEGF-B produced by microglia regulate the pathogenic activities of astrocytes in the experimental autoimmune encephalomyelitis (EAE) mouse model of multiple sclerosis. Microglia-derived TGF α acts via the ErbB1 receptor in astrocytes to limit their pathogenic activities and EAE development. Conversely, microglial VEGF-B triggers FLT-1 signalling in astrocytes and worsens EAE. VEGF-B and TGF α also participate in the microglial control of human astrocytes. Furthermore, expression of TGF α and VEGF-B in CD14⁺ cells correlates with the multiple sclerosis lesion stage. Finally,

metabolites of dietary tryptophan produced by the commensal flora control microglial activation and TGF α and VEGF-B production, modulating the transcriptional program of astrocytes and CNS inflammation through a mechanism mediated by the aryl hydrocarbon receptor. In summary, we identified positive and negative regulators that mediate the microglial control of astrocytes. Moreover, these findings define a pathway through which microbial metabolites limit pathogenic activities of microglia and astrocytes, and suppress CNS inflammation. This pathway may guide new therapies for multiple sclerosis and other neurological disorders.

Microglia are reported to express aryl hydrocarbon receptor (AHR)^{6,7}. To investigate the role of microglial AHR on CNS inflammation, we generated *Cx3cr1^{creERT2}Ahr^{fl/fl}* mice (CX3CR1-AHR mice) in which the *Cx3cr1* promoter drives the expression of Cre recombinase

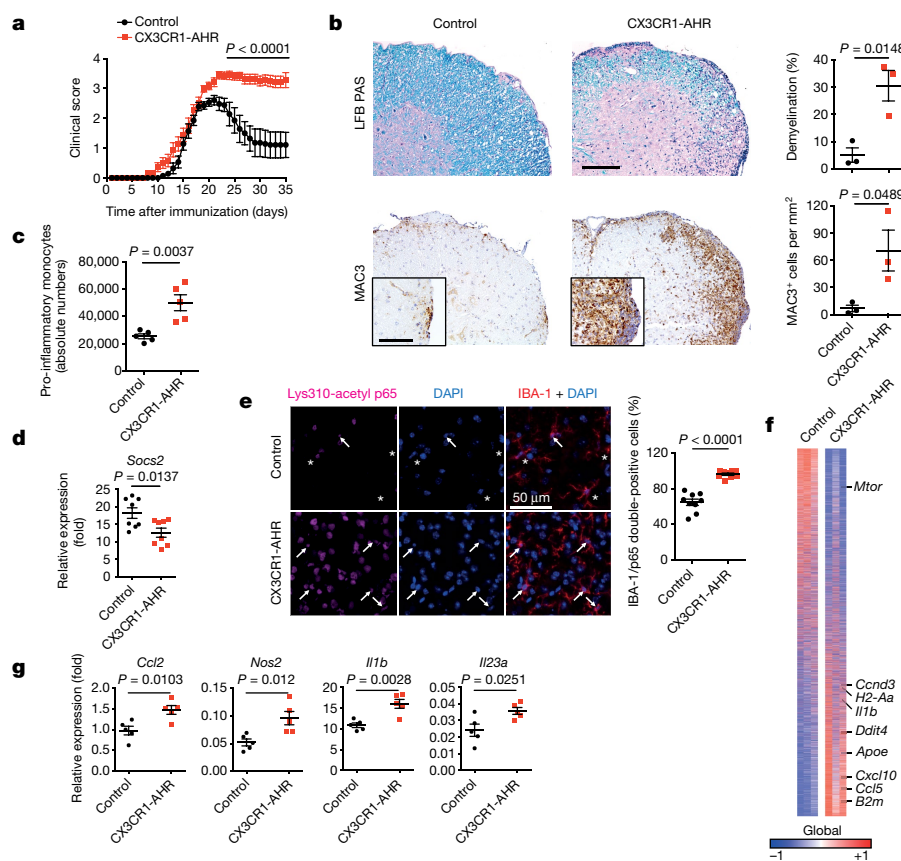


Fig. 1 | AHR limits microglial pro-inflammatory transcriptional responses during EAE. **a**, EAE clinical scores in control and CX3CR1-AHR mice ($n = 10$ mice per group). Data are representative of two independent experiments. **b**, Representative spinal cord sections from control and CX3CR1-AHR EAE mice stained for Luxol Fast blue (LFB) and periodic acid-Schiff (PAS) for demyelination (top), and MAC3 for macrophage infiltration (bottom). Representative of three sections from three mice. Right, quantification of demyelination and macrophage infiltration. **c**, Pro-inflammatory monocytes in the CNS of control and CX3CR1-AHR EAE mice. Data are representative of two independent experiments with $n = 5$ mice per group. **d**, Microglial mRNA expression determined by quantitative PCR (qPCR) in control ($n = 8$) and CX3CR1-AHR ($n = 8$) EAE mice. **e**, Left, Lys310-acetyl p65 in IBA-1-positive cells in control and CX3CR1-AHR EAE mice. Right, quantification of IBA-1/p65 double-positive cells. Data are representative of two independent experiments with $n = 9$ mice per group. **f**, Heat map of 9,957 genes expressed in microglia from control and CX3CR1-AHR mice ($n = 3$ mice per group). Gene expression is row-centred and log₂-transformed, and saturated at levels -0.5 and $+0.5$ for visualization satisfying a false discovery rate (FDR) < 0.1 . **g**, Microglial mRNA expression determined by qPCR in control ($n = 5$) and CX3CR1-AHR ($n = 5$) EAE mice. Data in **a–d**, **g** are mean \pm s.e.m. P values were determined by two-way analysis of variance (ANOVA) (**a**) or two-sided Student's t -test (**b–e**, **g**).

¹Ann Romney Center for Neurologic Diseases, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. ²Institute of Neuropathology, Medical Faculty, University of Freiburg, Freiburg, Germany. ³Neuroimmunology Unit, Montreal Neurological Institute, Department of Neurology and Neurosurgery, McGill University, Montreal, Quebec, Canada. ⁴Department of Pathobiology, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, PA, USA. ⁵INSERM, UMR 1064, Nantes, France. ⁶BIOSS Centre for Biological Signaling Studies, University of Freiburg, Freiburg, Germany. ⁷Broad Institute of MIT and Harvard, Cambridge, MA, USA. *e-mail: fquintana@rics.bwh.harvard.edu

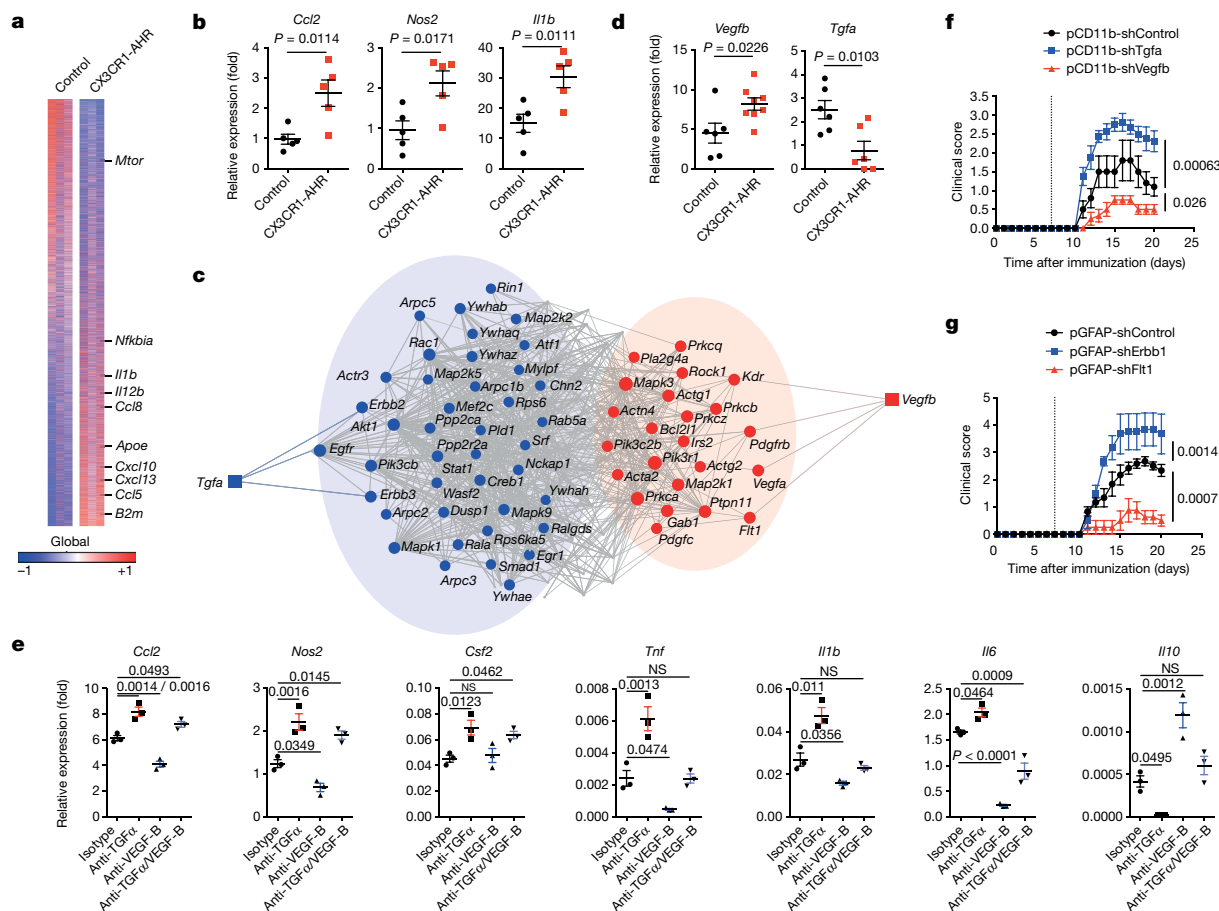


Fig. 2 | AHR-regulated microglial TGF α and VEGF-B control astrocytes during EAE. **a**, Heat map of 14,823 genes (detected at 0.1 level in at least two of three samples) expressed in astrocytes from control and CX3CR1-AHR mice. Gene expression is row-centred log₂-transformed and saturated at -0.5 and +0.5 levels for visualization satisfying an FDR < 0.1. *n* = 3 independent biological samples per group. **b**, mRNA expression determined by qPCR in control and CX3CR1-AHR EAE mice (*n* = 5 per group). **c**, Network diagram of differentially regulated genes in astrocytes and their predicted upstream regulators in microglia (*n* = 3 independent samples per group). **d**, Microglial mRNA expression determined by qPCR in control and CX3CR1-AHR EAE mice. Data are representative of two independent experiments with *n* = 6 control and

n = 8 (*Vegfb*) or *n* = 6 (*Tgfa*) CX3CR1-AHR mice per group.

e, Effect of MCM and blocking antibodies to TGF α and VEGF-B on gene expression in primary astrocytes determined by qPCR after 24 h. Data are representative of three independent experiments with three biological replicates. **f**, **g**, EAE in C57Bl/6J mice injected with lentiviral knockdown constructs targeting *Tgfa*, *Vegfb* or control in microglia (**f**), or *ErbB1* (also known as *Egfr*), *Flt1* or control in astrocytes (**g**). Data are representative of two independent experiments with *n* = 5 mice per group. Data in **b**, **e**–**g** are mean ± s.e.m. *P* values were determined by two-sided Student's *t*-test (**b**, **d**), one-way ANOVA followed by Tukey's post-hoc test (**e**) or two-way ANOVA (**f**, **g**). NS, not significant.

fused to an oestrogen ligand-binding domain⁸. After treatment of CX3CR1-AHR mice with tamoxifen, AHR-expressing peripheral CX3CR1⁺ cells are replenished from the bone marrow while microglia remain AHR deficient without impaired survival (Extended Data Fig. 1a–d). Microglial AHR deletion worsened EAE, increasing demyelination and CNS monocyte recruitment (Fig. 1a–c), the T-cell response was unaffected (Extended Data Fig. 1e, f). AHR deletion in peripheral CX3CR1⁺ cells achieved by chronic tamoxifen administration to bone marrow chimaeras of wild-type mice reconstituted with CX3CR1-AHR bone marrow⁹ led to earlier EAE onset, without affecting maximal scores and disease recovery (Extended Data Fig. 1g). AHR deficiency in both CNS-resident and peripheral CX3CR1⁺ cells accelerated EAE onset and impaired recovery (Extended Data Fig. 1h). Collectively, these data suggest that microglial AHR limits EAE.

NF- κ B controls microglial responses during EAE⁸, and AHR can limit NF- κ B activation in a SOCS2-dependent manner^{10,11}. The deletion of microglial AHR decreased *Socs2* expression and increased NF- κ B p65 nuclear localization in spinal cord Iba-1⁺ myeloid cells during EAE (Fig. 1d, e). Moreover, AHR deletion led to the upregulation of transcripts associated with microglial activation (*Apoe*, *Ddit4* and *B2m*), inflammation and neurodegeneration (*Ccl2*, *Nos2*, *Il1b* and *Il23a*)^{8,12} (Fig. 1f, g).

Microglia modulate astrocyte phenotype and function¹³. Indeed, microglial AHR deletion upregulated the expression of genes in astrocytes associated with inflammation and neurodegeneration, such as *Ccl2*, *Il1b* and *Nos2* (Fig. 2a, b). Bioinformatic analyses aimed to identify candidate cause and effect relationships between dysregulated transcriptional responses in microglia and astrocytes identified two transcriptional modules in astrocytes, potentially controlled by microglia-produced *Tgfa* and *Vegfb* during EAE (Extended Data Fig. 2a and Fig. 2c). Similar microglial *Vegfb* expression levels were detected throughout the CNS during EAE; *Tgfa* expression was slightly decreased in spinal cord microglia (Extended Data Fig. 2b).

Microglial AHR deletion decreased *Tgfa* and increased *Vegfb* expression during EAE (Fig. 2d). AHR regulates gene expression by direct interactions with target DNA regions, and also by controlling other transcription factors such as NF- κ B^{14,15}. We identified AHR and NF- κ B responsive elements (XREs and NREs, respectively) in the *Vegfb* and *Tgfa* promoters (Extended Data Fig. 2c). AHR deletion increased NF- κ B p65 recruitment to NREs in the *Vegfb* promoter in microglia during EAE (Extended Data Fig. 2d). NF- κ B p65 transactivated the *Vegfb* promoter in reporter assays; AHR suppressed this transactivation as well as *Vegfb* promoter basal activity (Extended Data Fig. 2e). AHR was also

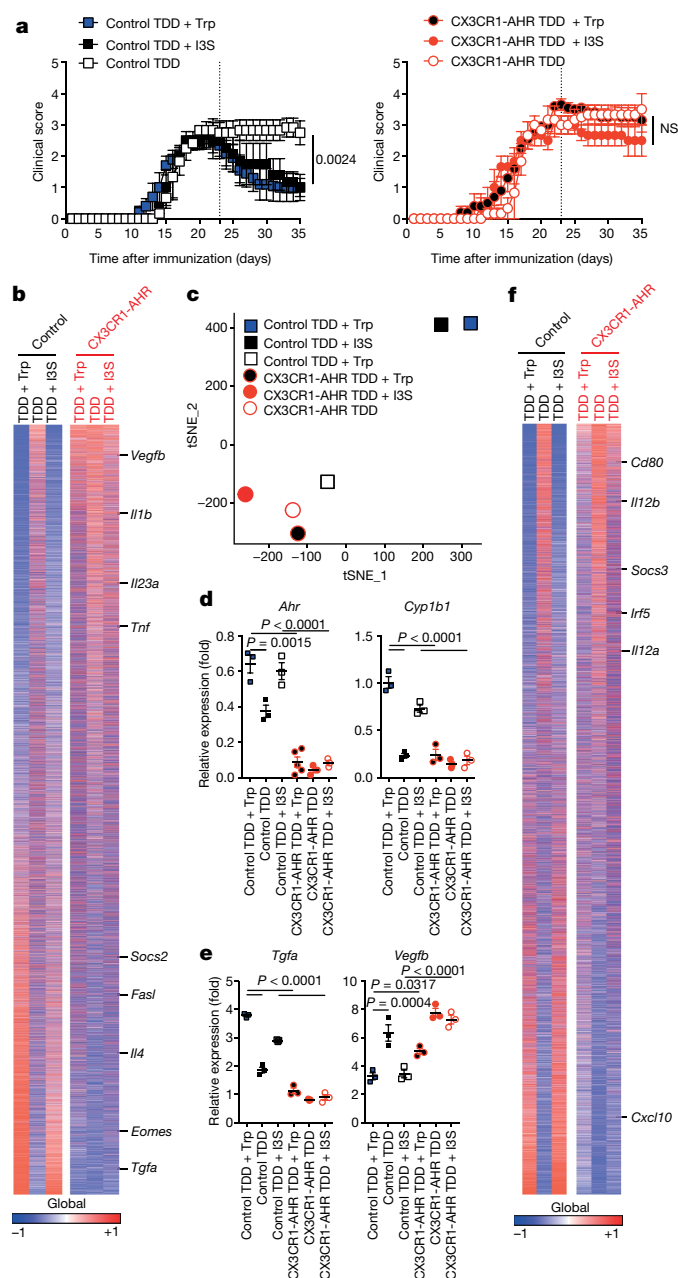


Fig. 3 | Trp metabolites control microglia-astrocyte interactions and CNS inflammation. **a**, Clinical scores in control (left) and CX3CR1-AHR (right) mice treated with TDD, TDD plus Trp or TDD plus I3S from day 21 after EAE induction ($n = 10$ mice per group). Data are representative of two independent experiments. **b**, Microglia were isolated and subjected to RNA sequencing (RNA-seq). Heat map of expressed genes of normalized reads of two independent samples per group. **c**, t -distributed stochastic neighbour embedding (tSNE) plot of RNA-seq data areolated from microglia of mice as in **b**. **d**, Microglial mRNA expression determined by qPCR in EAE mice as in **a**. Data are representative of two independent experiments with three replicates. **e**, mRNA expression determined by qPCR in microglia from EAE mice as in **a**. Data are representative of two independent experiments with three replicates. **f**, Heat map depicting mRNA expression in astrocytes from EAE mice as in **a**, as determined by RNA-seq of normalized reads of two independent samples per group. Data in **a**, **d** and **e** are mean \pm s.e.m. P values were derived by two-way ANOVA (**a**), one-way ANOVA followed by Tukey's post-hoc test (**d**, **e**).

recruited to XREs in the *Tgfa* promoter in microglia (Extended Data Fig. 2f) and transactivated the *Tgfa* promoter in reporter studies (Extended Data Fig. 2g). These findings suggest that AHR regulates microglial *Tgfa* and *Vegfb* expression through its direct effects on

the *Tgfa* and *Vegfb* promoters, and through its ability to limit NF- κ B activation.

We then analysed the effects of microglial TGF α and VEGF-B on astrocytes. Microglial AHR inhibition decreased *Tgfa* and increased *Vegfb* expression (Extended Data Fig. 3a), and it boosted pro-inflammatory *Ccl2* and neurotoxic *Nos2* expression induced in astrocytes by microglial supernatants (Extended Data Fig. 3b–f). Antibody blockade showed that TGF α and VEGF-B mediated these effects with a relative dominance of TGF α suppressive effects (Fig. 2e). Recombinant TGF α decreased pro-inflammatory chemokine (*Ccl2* and *Csf2*) and cytokine (*Il6*) expression induced in mouse astrocytes by TNF and IL-1 β , whereas it enhanced *Il10* expression (Extended Data Fig. 3g). Conversely, VEGF-B boosted *Ccl2*, *Csf2* and *Nos2* expression in astrocytes. Similarly, VEGF-B pretreatment enhanced the toxicity of astrocyte-conditioned medium (ACM) towards neurons and oligodendrocytes; TGF α reduced this toxicity (Extended Data Fig. 3h,i). VEGF-B pretreatment also enhanced pro-inflammatory monocyte recruitment and microglia activation by ACM; these activities were inhibited by TGF α (Extended Data Fig. 3j,k). These data suggest that TGF α and VEGF-B control astrocyte functions that contribute to CNS pathology.

To investigate the interaction between microglia and astrocytes in vivo, we knocked down *Tgfa* and *Vegfb* expression in microglia using lentivirus-delivered short-hairpin RNAs (shRNAs) expressed under the control of the *Itgam* (encoding CD11b) promoter. The knockdown did not affect astrocyte numbers or morphology, nor *Tgfa* and *Vegfb* expression in CNS-infiltrating monocytes (Extended Data Fig. 4a–f). Microglial *Tgfa* knockdown worsened EAE, whereas *Vegfb* knockdown ameliorated it (Fig. 2f). Similar observations were made when the TGF α or VEGF-B receptors ErbB1 or FLT-1, respectively, were knocked down in astrocytes (Fig. 2g, Extended Data Fig. 4a, b, e, g). Of note, VEGF-B administration did not induce demyelination in naive mice (Extended Data Fig. 4h), suggesting that VEGF-B synergizes with other factors to boost EAE pathology. Moreover, the knockdown of *Tgfa* and *Vegfb* in astrocytes or of their receptors in microglia did not affect EAE (Extended Data Fig. 5), supporting a microglia to astrocyte directionality in their effects.

Transcriptional analyses suggested that TGF α –ErbB1 and VEGF-B–FLT-1 regulate NF- κ B in astrocytes (Extended Data Fig. 6a–d), known to drive their pathogenic activities during CNS inflammation^{10,16–18}. Indeed, NF- κ B signalling in astrocytes was increased after microglial AHR deletion (Extended Data 2a). Interestingly, VEGF-B boosted astrocytic NF- κ B activation; this boost was inhibited by TGF α (Extended Data Fig. 6e). Moreover, NF- κ B blockade suppressed the increase in pro-inflammatory gene expression induced by VEGF-B in astrocytes (Extended Data Fig. 6f, g). Collectively, these findings suggest that by controlling NF- κ B signalling, VEGF-B and TGF α modulate astrocyte pathogenic activities.

The microbial metabolism of dietary tryptophan (Trp) generates AHR agonists such as I3S, which limits astrocyte pathogenic activities and EAE development^{10,19,20}. To investigate the role of microglial AHR in the control of CNS inflammation by dietary Trp metabolites, we subjected control and CX3CR1-AHR mice to a Trp-depleted diet (TDD) initiated 21 days after EAE induction. Trp depletion interfered with disease recovery in control mice (Fig. 3a). Trp or I3S administration ameliorated EAE in control but not in CX3CR1-AHR TDD-fed mice (Fig. 3a), suggesting that microglial AHR participates in EAE amelioration by Trp metabolites. In addition, a TDD initiated 14 days after EAE induction worsened disease in control mice and also in mice with AHR-deficient microglia or astrocytes (Extended Data Fig. 7a–c), suggesting that Trp metabolites limit CNS inflammation via both microglial and astrocytic AHR. Indeed, I3S administration initiated 14 days after EAE induction ameliorated disease via AHR in astrocytes and microglia (Extended Data Fig. 7b, c). Similar results were obtained when AHR was knocked down in astrocytes or microglia (Extended Data Fig. 7d–g). Collectively, these findings suggest that microglial AHR deletion

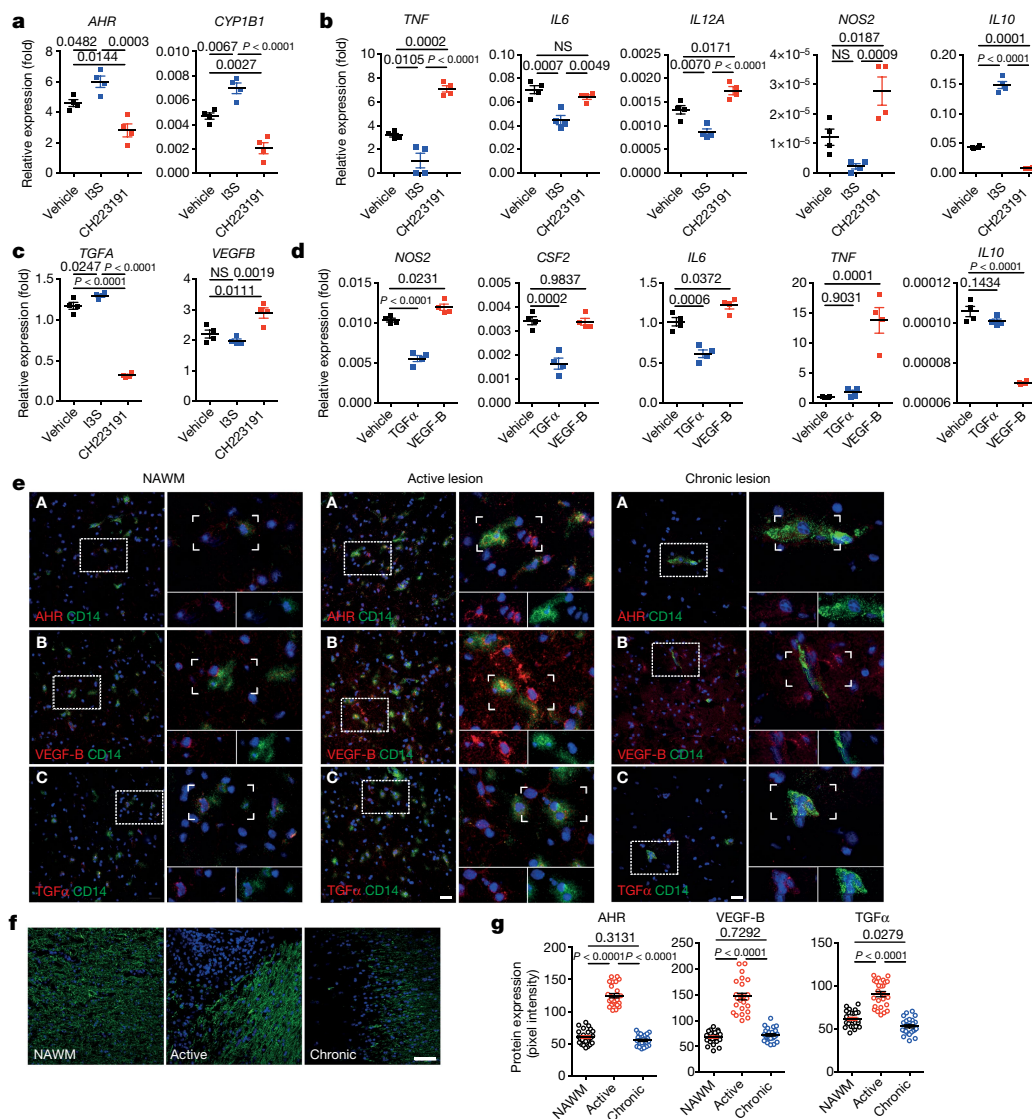


Fig. 4 | VEGF-B and TGF α control human astrocytes and are expressed by CD14⁺ cells in MS lesions. **a–c**, mRNA expression determined by qPCR in human microglia activated in the presence of I3S or the AHR antagonist CH223191 24 h after activation. $n = 4$ biological replicates. Data are mean \pm s.e.m. and representative of three independent experiments. P values were derived by one-way ANOVA followed by Tukey's post-hoc test. **d**, mRNA expression determined by qPCR in primary human astrocytes activated in the presence of TGF α or VEGF-B. Data are mean \pm s.e.m. and representative of three independent experiments from four biological replicates. P values were derived by one-way ANOVA followed by Tukey's post-hoc test. **e**, Immunofluorescence staining for AHR (A, red), VEGF-B (B, red), or TGF α (C, red), CD14 (green),

and DAPI (blue) in human brain samples corresponding to normally appearing white matter (NAWM), active, and chronic MS lesions. Data are representative of $n = 12$ fields from three distinct MS brains. Insets highlight co-expression of AHR and CD14, VEGF-B and CD14, and TGF α and CD14. **f**, Myelin oligodendrocyte glycoprotein (green) staining in tissues from patients with MS. Nuclear staining was done using Hoechst (blue). Representative sections of NAWM, active and chronic lesions from patients with MS ($n = 3$). **g**, Quantification of AHR, VEGF-B and TGF α expression in NAWM, active, or chronic lesions in MS tissue. Data are mean \pm s.e.m. and representative of 25 fields from 3 distinct MS brains. P values were determined by one-way ANOVA followed by Tukey's post-hoc test.

renders astrocytes unresponsive to the anti-inflammatory effects of AHR ligands at later EAE stages.

The transcriptional response of microglia in TDD-fed control mice resembled that of AHR-deficient microglia (Fig. 3b, c). Indeed, the microglial expression of *Ahr* and its target gene *Cyp1b1* was suppressed in TDD-fed and CX3CR1-AHR mice, and could be restored by Trp or I3S supplementation in control but not in CX3CR1-AHR mice (Fig. 3d). Moreover, dietary Trp metabolites promoted *Socs2* expression, which mediates NF- κ B regulation by AHR^{10,11} and suppressed the microglial expression of NF- κ B dependent transcripts such as *Tnf* (also known as *Tnfa*) in an AHR-dependent manner (Fig. 3b and Extended Fig. Data 8a). In addition, dietary Trp metabolites also regulated microglial *Tgfa* and *Vegfb* expression via AHR (Fig. 3e). Accordingly, astrocytes from CX3CR1-AHR and

TDD-treated control mice showed increased expression of genes linked to EAE pathogenesis such as *Ccl2* and *Nos2* (Extended Data Fig. 8b and Fig. 3f). Collectively, these findings suggest that dietary Trp metabolites such as I3S limit NF- κ B driven pro-inflammatory programs in microglia and suppress their ability to promote pro-inflammatory activities in astrocytes.

We validated our observations using human samples. AHR was activated by I3S and inhibited by the antagonist CH223191 in primary human microglia, as indicated by the expression of *AHR* and its target *CYP1B1* (Fig. 4a). Microglial AHR activation suppressed pro-inflammatory and neurotoxic gene expression (*TNF*, *IL6*, *IL12A* and *NOS2*) and boosted anti-inflammatory *IL10* expression; AHR activation also promoted *TGFA* and suppressed *VEGFB* expression in human microglia (Fig. 4b, c). More importantly, TGF α and VEGF-B suppressed and

boosted, respectively, pro-inflammatory gene expression in primary human astrocytes (Fig. 4d).

Finally, we analysed AHR, TGF α and VEGF-B expression on brain samples from patients with MS. We detected AHR, TGF α and VEGF-B expression in CD14⁺ cells (microglia and recruited monocytes) in the normally appearing white matter (NAWM), demyelinated active and chronic multiple sclerosis (MS) lesions; the highest AHR, VEGF-B and TGF α expression was detected in CD14⁺ cells in MS active lesions (Fig. 4e, f and Extended Data Fig. 8c). VEGF-B and AHR expression in chronic MS lesions were comparable to their levels in NAWM, whereas TGF α expression was decreased to levels below those detected in the NAWM, resulting in a higher VEGF-B/TGF α ratio than in NAWM (Fig. 4g and Extended Data Fig. 8d). These findings suggest that VEGF-B and TGF α participate in the control of astrocytes by microglia in humans and contribute to MS pathogenesis.

In summary, we found that AHR-controlled microglial VEGF-B and TGF α regulate astrocyte pathogenic activities during EAE. VEGF-A promotes CNS pathology by several mechanisms including angiogenesis induction²¹, and microglia²² and T cell²³ stimulation, but less is known about VEGF-B, which does not promote angiogenesis in the CNS²⁴ and shows neuroprotective effects in some models²⁵. Our data suggest that FLT-1 activation in astrocytes by VEGF-B produced by microglia and other sources²⁶ promotes CNS inflammation, identifying VEGF-B–FLT-1 signalling inhibitors as candidate therapeutics for CNS inflammation. Conversely, TGF α induces astrogliosis and neuroprotective factor production, and increases neuronal survival and axonal growth in multiple contexts, including models of spinal cord injury^{27,28}. Indeed, based on the promotion of axon regeneration by reactive astrocytes in spinal cord injury models²⁹, it is tempting to speculate that microglial TGF α promotes these beneficial astrocyte activities. Future studies should address whether the control of TGF α –ErbB1 signalling via AHR contributes to the beneficial effects of commensal bacteria on spinal cord injury³⁰. In conclusion, our findings define a gut–brain axis by which metabolites of dietary Trp controlled by the commensal flora act directly on CNS-resident microglia and astrocytes¹⁰ to limit inflammation and neurodegeneration via AHR.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0119-x>.

Received: 14 July 2017; Accepted: 11 April 2018;

Published online: 16 May 2018

- Ben Haim, L. & Rowitch, D. H. Functional diversity of astrocytes in neural circuit regulation. *Nat. Rev. Neurosci.* **18**, 31–41 (2017).
- Khakh, B. S. & Sofroniew, M. V. Diversity of astrocyte functions and phenotypes in neural circuits. *Nat. Neurosci.* **18**, 942–952 (2015).
- Sofroniew, M. V. Astrocyte barriers to neurotoxic inflammation. *Nat. Rev. Neurosci.* **16**, 249–263 (2015).
- Liddel, S. A. & Barres, B. A. Reactive astrocytes: production, function, and therapeutic potential. *Immunity* **46**, 957–967 (2017).
- Prinz, M. & Priller, J. Microglia and brain macrophages in the molecular age: from origin to neuropsychiatric disease. *Nat. Rev. Neurosci.* **15**, 300–312 (2014).
- Buttgereit, A. et al. Sall1 is a transcriptional regulator defining microglia identity and function. *Nat. Immunol.* **17**, 1397–1406 (2016).
- Lee, Y. H. et al. Aryl hydrocarbon receptor mediates both proinflammatory and anti-inflammatory effects in lipopolysaccharide-activated microglia. *Glia* **63**, 1138–1154 (2015).
- Goldmann, T. et al. A new type of microglia gene targeting shows TAK1 to be pivotal in CNS autoimmune inflammation. *Nat. Neurosci.* **16**, 1618–1626 (2013).
- Croxford, A. L. et al. The cytokine GM-CSF drives the inflammatory signature of CCR2⁺ monocytes and licenses autoimmunity. *Immunity* **43**, 502–514 (2015).
- Rothhammer, V. et al. Type I interferons and microbial metabolites of tryptophan modulate astrocyte activity and central nervous system inflammation via the aryl hydrocarbon receptor. *Nat. Med.* **22**, 586–597 (2016).
- Yeste, A. et al. Tolerogenic nanoparticles inhibit T cell-mediated autoimmunity through SOCS2. *Sci. Signal.* **9**, ra61 (2016).

- Matcovitch-Natan, O. et al. Microglia development follows a stepwise program to regulate brain homeostasis. *Science* **353**, aad8670 (2016).
- Liddel, S. A. et al. Neurotoxic reactive astrocytes are induced by activated microglia. *Nature* **541**, 481–487 (2017).
- Quintana, F. J. & Sherr, D. H. Aryl hydrocarbon receptor control of adaptive immunity. *Pharmacol. Rev.* **65**, 1148–1161 (2013).
- Stockinger, B., Di Meglio, P., Gialitakis, M. & Duarte, J. H. The aryl hydrocarbon receptor: multitasking in the immune system. *Annu. Rev. Immunol.* **32**, 403–432 (2014).
- Mayo, L. et al. Regulation of astrocyte activation by glycolipids drives chronic CNS inflammation. *Nat. Med.* **20**, 1147–1156 (2014).
- Rothhammer, V. et al. Sphingosine 1-phosphate receptor modulation suppresses pathogenic astrocyte activation and chronic progressive CNS inflammation. *Proc. Natl Acad. Sci. USA* **114**, 2012–2017 (2017).
- Wheeler, M. A. & Quintana, F. J. Regulation of Astrocyte functions in multiple sclerosis. *Cold Spring Harb. Perspect. Med.* <https://doi.org/10.1101/cshperspect.a029009> (2018).
- Gutiérrez-Vázquez, C. & Quintana, F. J. Regulation of the immune response by the aryl hydrocarbon receptor. *Immunity* **48**, 19–33 (2018).
- Wikoff, W. R. et al. Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proc. Natl Acad. Sci. USA* **106**, 3698–3703 (2009).
- Girolamo, F., Coppola, C., Ribatti, D. & Trojano, M. Angiogenesis in multiple sclerosis and experimental autoimmune encephalomyelitis. *Acta Neuropathol. Commun.* **2**, 84 (2014).
- Mosher, K. I. et al. Neural progenitor cells regulate microglia functions and activity. *Nat. Neurosci.* **15**, 1485–1487 (2012).
- Mor, F., Quintana, F. J. & Cohen, I. R. Angiogenesis-inflammation cross-talk: vascular endothelial growth factor is secreted by activated T cells and induces Th1 polarization. *J. Immunol.* **172**, 4618–4623 (2004).
- Gaal, E. I. et al. Comparison of vascular growth factors in the murine brain reveals placenta growth factor as prime candidate for CNS revascularization. *Blood* **122**, 658–665 (2013).
- Li, X., Kumar, A., Zhang, F., Lee, C. & Tang, Z. Complicated life, complicated VEGF-B. *Trends Mol. Med.* **18**, 119–127 (2012).
- Nag, S., Eskandarian, M. R., Davis, J. & Eubanks, J. H. Differential expression of vascular endothelial growth factor-A (VEGF-A) and VEGF-B after brain injury. *J. Neuropathol. Exp. Neurol.* **61**, 778–788 (2002).
- Junier, M. P. What role(s) for TGF α in the central nervous system? *Prog. Neurobiol.* **62**, 443–473 (2000).
- White, R. E., Yin, F. Q. & Jakeman, L. B. TGF- α increases astrocyte invasion and promotes axonal growth into the lesion following spinal cord injury in mice. *Exp. Neurol.* **214**, 10–24 (2008).
- Anderson, M. A. et al. Astrocyte scar formation aids central nervous system axon regeneration. *Nature* **532**, 195–200 (2016).
- Kigerl, K. A. et al. Gut dysbiosis impairs recovery after spinal cord injury. *J. Exp. Med.* **213**, 2603–2620 (2016).

Acknowledgements This work was supported by grants NS087867, ES02530, AI126880 and AI093903 from the National Institutes of Health, RSG-14-198-01-LIB from the American Cancer Society and RG4111A1 and JF2161-A-5 from the National Multiple Sclerosis Society to F.J.Q. F.J.Q. and J.A. received support from International Progressive Multiple Sclerosis Alliance grant PA-1604-08459. V.R. received support from an educational grant from Mallinkrodt Pharmaceuticals (A219074) and by a fellowship from the German Research Foundation (DFG RO4866 1/1). M.P. is supported by the BMBF-funded competence network of multiple sclerosis (KKNMS), the Sobek-Stiftung and the DFG (SFB 992, SFB1140, SFB/TRR167, Reinhart-Koselleck-Grant) and the Ministry of Science, Research and the Arts, Baden-Wuerttemberg (Sonderlinie ‘Neuroinflammation’). Human fetal tissue came from the Human Fetal Tissue Repository (Albert Einstein College of Medicine) and from the University of Washington Birth Defects Research Laboratory (BDRL, Project Number 5R24HD000836-51).

Reviewer information Nature thanks S. Liddel, M. Platten, H. Wekerle and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions V.R., D.M.B., M.C.T., C.-C.C., A.A.-F., K.A.d.L., C.G.-V., P.H., O.S., M.B.I., L.H., T.N., M.Bo., M.W., L.L.D., D.A.L. and J.I.A. performed in vitro and in vivo experiments, J.A. and M.P. provided unique reagents, discussed and/or interpreted findings, E.C.T. performed bioinformatics, V.R. and F.J.Q. wrote the manuscript and F.J.Q. designed and supervised the study and edited the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0119-x>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0119-x>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to F.J.Q.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Animals. C57BL/6J mice were obtained from the Jackson Laboratory and were all female. *Cx3cr1^{creERT2}* mice⁸ were a gift from S. Jung and were bred to *Ahr^{fl/fl}* mice. To delete microglial AHR, 4–5-week-old mice were injected subcutaneously with 4 mg tamoxifen (Sigma) in 200 µl warm corn oil at two time points 48 h apart. *Cx3cr1^{creERT2}*-negative *Ahr^{fl/fl}* mice were used as controls. Four weeks later, EAE was induced. To delete AHR in all CX3CR1-expressing cells, control and CX3CR1-AHR mice were gavaged weekly with tamoxifen starting from 5 weeks of age. EAE was induced and weekly tamoxifen gavages were continued after EAE induction.

Bone marrow chimaera were generated as previously described to minimize irradiation-induced artefacts^{31,32}. In brief, 7-week-old wild-type recipient mice were lethally irradiated with a dose of 9.5 Gy. One day later, mice were administered 5×10^6 bone marrow cells isolated from donor femora and tibiae by intravenous injection. Donors were *Cx3cr1^{cre}*-negative *Ahr^{fl/fl}* and *Cx3cr1^{cre}*-positive *Ahr^{fl/fl}* mice. Bone marrow recipients were then rested for 3 weeks and thereafter treated with weekly tamoxifen gavages for another 3 weeks; after a total of 6 weeks, EAE was induced as described below. Tamoxifen administration was continued weekly during EAE. All mice were on the C57BL/6 background and were kept in a pathogen-free facility at the Harvard Institutes of Medicine. All experiments were carried out in accordance with guidelines prescribed by the Institutional Animal Care and Use Committee (IACUC) at Harvard Medical School.

EAE induction and treatment. EAE was induced in 8-week-old mice by subcutaneous immunization with 150 µg myelin oligodendrocyte glycoprotein (MOG_{35–55}) peptide (Genemed Synthesis Inc.) emulsified in complete Freund's adjuvant (CFA, Difco Laboratories) per mouse, followed by administration of 200 ng pertussis toxin (PTX, List Biological Laboratories, Inc.) on days 0 and 2 as described¹⁰. Clinical signs of EAE were assessed as follows: 0, no signs of disease; 1, loss of tone in the tail; 2, hind limb paresis; 3, hind limb paralysis; 4, tetraplegia; 5, moribund. All agents were purchased from Sigma-Aldrich.

Isolation of cells from adult mouse CNS. Mononuclear cells were isolated from the CNS as previously described, and astrocytes, monocytes, and microglia were sorted as described before^{10,16,17}. Isolated CNS cells were stained with fluorochrome-conjugated antibodies to CD11b (M1/70, 1:100), CD45 (90, 1:100), Ly6C1 (HK1.4, 1:100), CD105 (N418, 1:100), CD140a (APA5, 1:100), CD11c (N418, 1:100), F4/80 (BM8, 1:50), O4 (O4, Miltenyi Biotec, 1:10), and CD19 (eBio1D3, 1:100). All antibodies were from eBioscience or BD Pharmingen, unless otherwise mentioned. Microglia were sorted as CD11b⁺ cells with low CD45 expression and low Ly6C1 (CD11b⁺CD45^{lo}Ly6C1^{lo}), inflammatory monocytes were considered as CD45^{hi}CD11b⁺Ly6C1^{hi}. Astrocytes were sorted as CD11b^{lo}CD45^{lo}Ly6C1^{lo}CD105^{lo}CD140a^{lo}CD11b^{lo}F4/80^{lo}O4^{lo}CD19^{lo} after the exclusion of lymphocytes, microglia, oligodendrocytes, and monocytes.

Flow cytometry staining and acquisition. Mononuclear cell suspensions were prepared as previously described^{10,16,17}. Antibodies for flow cytometry were purchased from eBioscience or BD Pharmingen and used at a concentration of 1:100 unless recommended otherwise by the manufacturer. Mouse AHR antibody (IC6697G) and mouse FLT-1 antibody (FAB4711A) were from R&D Systems, VEGF-B (RM0008-6E72) and TGFα (MF9) from Novus Biologicals, EGFR (D38B1) and p-p65 (93H1) from Cell Signaling. Cells were then analysed on a LSRII or MACSQuant flow cytometer (BD Biosciences and Miltenyi Biotec, respectively). As outlined in the individual figures, T-helper 1 (T_H1) cells were defined as CD3⁺CD4⁺IFNγ⁺IL-17⁺IL-10⁺FOXP3⁺, T_H17 cells as CD3⁺CD4⁺IFNγ⁺IL-17⁺IL-10⁺FOXP3⁺, T-regulatory (T_{reg}) cells as CD3⁺CD4⁺IFNγ⁺IL-17⁺FOXP3⁺, microglia as CD11b⁺CD45^{lo}Ly6C^{lo}, and pro-inflammatory monocytes as CD45^{hi}CD11b⁺Ly6C^{hi}.

RNA-seq. Mice were euthanized at day 25 after disease induction and astrocytes isolated as described above. RNA was sequenced using the strand-specific TruSeq protocol. High coverage (>50 M) strand-specific paired-end 76-base-pair (bp) reads were aligned to the mm10/GRCm38 mouse reference genome using TopHat v2.0.11³³. Gene expression levels were estimated for 38922 GenCode Release M2 (GRCm38.p2) mouse gene annotations using Cuffquant and Cuffnorm v2.2.1 quartile normalized FPKMs³³.

nCounter gene expression. In total, 50 ng of total RNA was hybridized with reporter and capture probes in custom-made astrocyte-targeted nCounter Gene Expression code set according to manufacturer's instructions (NanoString Technologies). Data were analysed using nSolver Analysis software.

qPCR. RNA was extracted using the RNeasy kit (Qiagen), cDNA was prepared and used for qPCR with the results normalized to *Gapdh*. All primers and probes were from Applied Biosystems. Mouse: *Ahr* Mm00478932_m1, *Aldh1a1* Mm00657317_m1, *Aqp4* Mm00802131_m1, *Ccl20* Mm01268754_m1, *Ccl2* Mm00441242_m1, *Ccl8* Mm01297183_m1, *Cxcl3* Mm01701838_m1, *Cyp1b1* Mm00487229_m1, *Erbp1* Mm01187858_m1, *Flt1* Mm00438980_m1, *Gapdh* Mm99999915_g1, *Gfap* Mm01253033_m1, *Il10* Mm00439614_m1, *Il12a* Mm00434165_m1, *Il23a* Mm01160011_g1, *Il6* Mm00446190_m1, *Itgam* Mm00434455_m1, *Nos2* Mm00440502_m1, *Tgfa* Mm00446232_m1, *Tnf*

Mm00443258_m1 and *Vegfb* Mm00442102_m1. Human: *AHR* Hs00169233_m1, *CCL2* Hs00234140_m1, *CYP1B1* Hs02382916_s1, *ERBB1* (also known as *EGFR*), *FLT1*, *IL6* Hs00985639_m1, *NOS2* Hs01075529_m1, *TGFA* Hs00608187_m1 and *TNF* Hs01113624_g1, *VEGFB* Hs00173634_m1.

T cell proliferation. Splenocytes were cultured in X-VIVO 15 medium (Lonza) and were plated for 72 h at a density of 5×10^5 cells per well with increasing concentrations of MOG_{35–55} peptide. During the final 16 h, cells were pulsed with 1 Ci [³H]thymidine (PerkinElmer), followed by collection on glass fibre filters and analysis of incorporated [³H]thymidine in a beta-counter (1450 MicroBeta TriLux; PerkinElmer). For intracellular cytokine staining, cells were stimulated for 4 h with PMA (phorbol 12-myristate 13-acetate; 50 ng ml⁻¹; Sigma), ionomycin (1 µg ml⁻¹; Sigma) and monensin (GolgiStop; 2 µM BD Biosciences). After staining of surface markers, cells were fixed and made permeable according to the manufacturer's instructions (BD Cytofix/Cytoperm Kit (BD Biosciences) or FOXP3 Fixation/Permeabilization (eBioscience).

Primary astrocyte and microglia cultures. Cerebral cortices from neonatal mice (1–3 days) were dissected, carefully stripped of their meninges, digested with 0.25% trypsin-EDTA and DNase I (1 mg ml⁻¹) for 15 min, and dispersed to single-cell level by passing through a cell strainer (70 µm). The cell suspension was then cultured at 37 °C in humidified 5% CO₂, 95% air on poly-L-lysine (Sigma) precoated 75 cm² cell culture flasks. Medium was replaced every 4–5 days. After 7–10 days, cells reached confluence and astrocytes and microglia were isolated by mild trypsinization with Trypsin-EDTA (0.06%) as previously described^{10,16,17}. Cells were >95% astrocytes as determined by staining with GFAP or GLAST, with less than 5% contamination of CD11b⁺ microglia cells (not shown). Conversely, microglia cultures stained CD11b⁺CD45^{lo}Ly6C1^{lo} >95%. After the isolation procedure, cells were further plated as required for the specific experiments. Concentrations of agents were 100 ng ml⁻¹ for LPS (Sigma), 50 µg ml⁻¹ for poly(I:C), 100 ng ml⁻¹ for IL-1β, 50 ng ml⁻¹ for TNE, 0.1 ng ml⁻¹ TGFα, 10 ng ml⁻¹ VEGF-B (all R&D Systems), 50 µg ml⁻¹ 3-indoxyl-sulfate (Sigma), 100 nM NF-κB Blocker Bengamide B (Tocris). Unless otherwise indicated, RNA was isolated 24 h after start of treatment. For western blot analysis, cells were pretreated with I3S or vehicle for 24 h, thereafter LPS was added and protein prepared after 2 h.

Plasmids. Constructs encoding p65, AHR, pTgfa-Luc, pVegfb-Luc, as well as pTK-Renilla were obtained from Addgene. The pLenti-GFAP-EGFP-mir30-shAct1 vector was a gift from G.-X. Zhang³⁴. The pLenti-CD11b-EGFP-mir30-shRNA was also provided by G.-X. Zhang, who generated it by exchanging the *Gfap* promoter with the *Cd11b* (also known as *Itgam*) promoter.

In vivo knockdown with shRNA lentivirus. shRNA sequences against *Ahr*, *Tgfa*, *Vegfb*, *Erbp1* or *Flt1* and a non-targeting control shRNA were cloned into pLenti-GFAP-EGFP-mir30-shRNA or pLenti-CD11b-EGFP-mir30-shRNA using the following validated shRNA sequence against *Ahr* (5'-CCGGCATCGACATAACGGACGAAATCTCGAGATTTCGTCCGTTATGTCGATGTTT TTTG-3'), *Tgfa* (5'-CCGGTCTCGTCAGGATGCGTGTCTTATCTCGAGATAAGACACGCATCCGTACGATTTTTTG-3'), *Vegfb* (5'-CCGGGCGCAATGTGAATGCAGACCAACCTCGAGTTG GTCTGCATTCACATTCGCTTTTGTG-3'), *Erbp1* (5'-CCGGGCTGGATGATAGATGCTGATACTCGAGTATCAGCATCTATCATCCAGCTTTTTTG-3') or *Flt1* (5'-CCGGCGTGACCTTAAATCGTGCTTTCTCGAGAAAGCACGATTAAGGTCACGTTTTT-3') as described¹⁶. Lentivirus particles were generated by transfecting HEK-293T cells (Invitrogen) with the pLenti-GFAP-EGFP-mir30-shRNA or pLenti-CD11b-EGFP-mir30-shRNA vector and the ViraPower Packaging mix (helper plasmids pLP1, pLP2, pLP/VSV-G, Invitrogen). Supernatants were collected, filtered through a 0.45-µm PVDF filter, and concentrated overnight with the Lenti-X concentrator kit (Clontech). The viral titre was determined using the Lenti-X qRT-PCR titration kit (Clontech). For in vivo knockdown, immunized mice were anaesthetized at indicated time points, positioned in a Kopf Stereotaxic Alignment System and injected with 10⁷ IU of respective virus using a Hamilton syringe 0.44 mm posterior to the bregma, 1.0 mm lateral to it and 2.3 mm below the skull surface. The injection system was retracted slowly, skin incisions closed carefully by surgical sutures, mice allowed to wake up in a cage pre-warmed with a red light and mice checked twice daily thereafter.

Assessment of toxicity towards neurons and oligodendrocytes. N2A neuronal cells (ATCC CCL-131, ATCC) or mouse oligodendrocytes (Celprogen, 11004-02) were grown in 96-well plates and pre-activated with mouse IFNγ (100 ng ml⁻¹, R&D Systems) for 24 h. Thereafter, medium was supplemented after extensive PBS washes with ACM. Cytotoxicity was determined by quantifying LDH release (CytoTox 96 Non-Radioactive Cytotoxicity Assay, Promega) after 24 h as suggested by manufacturer's protocol.

Monocyte migration assay. Splenic monocytes from wild-type mice were pre-enriched by CD11b beads (Miltenyi) and sorted for F4/80⁺SSC^{lo}Ly6C^{hi}. These monocytes were seeded in the upper chamber of a 24-well cell culture insert with 5 µm pore-size (Corning) containing ACM. Migrating monocytes were quantified in the lower chamber after 3 h.

Microglia polarization assays. Wild-type microglia were co-cultured with astrocytes pre-treated with TGF α or VEGF-B and extensively washed. After 24 h, microglia were re-isolated, RNA was isolated, transcribed and subjected to qPCR analysis.

Subcellular fractionation and immunoblot analysis. In vitro microglia cultures were treated as indicated in specific experiments, subcellular fractions generated using Cell Fractionation kit (Cell Signaling) and 10 μ g of nuclear and cytoplasmic fractions were separated by 4–12% Bis-Tris Nupage gels (Invitrogen) and transferred onto PVDF membranes (Millipore). As primary antibodies rabbit anti-GAPDH monoclonal antibody (14C10, Cell Signaling), anti-histone H3 rabbit polyclonal antibody (EMD Millipore), anti-NF- κ B p65 rabbit monoclonal antibody (D14E12, Cell Signaling) were used, followed by goat anti-rabbit IgG horseradish peroxidase (HRP)-linked antibody (7074S, Cell Signaling). All antibodies were used at a dilution of 1:1,000. Blots were developed using the SuperSignal West Femto Maximum sensitivity kit (Thermo Scientific/Life Technologies). Data quantification was done using Image J software 1.48v (NIH) and specific signals normalized to GAPDH (cytoplasm) or histone 3 (nucleus).

Chromatin immunoprecipitation. Cells were cross-linked with 1% paraformaldehyde and lysed with 350 μ l lysis buffer (1% SDS, 10 mM EDTA, 50 mM Tris-HCl, pH 8.1) containing 1 \times protease inhibitor cocktail (Roche Molecular Biochemicals). Chromatin was sheared by sonication and supernatants were collected after centrifugation and diluted in chromatin immunoprecipitation incubation buffer (1% Triton X-100, 2 mM EDTA, 150 mM NaCl, 20 mM Tris-HCl, pH 8.0). Ten micrograms of antibody was prebound for 6 h to protein A- and protein G-Dynal magnetic beads (Invitrogen) and washed three times with PBS plus 1% bovine serum albumin (BSA), and then added to the diluted chromatin and immunoprecipitated rotating overnight. The magnetic bead–chromatin complexes were then washed three times in RIPA buffer (50 mM HEPES (pH 7.6), 1 mM EDTA, 0.7% sodium deoxycholate, 1% NP-40, 0.5 M LiCl) and then twice with Tris-EDTA (TE) buffer. Immunoprecipitated chromatin was then extracted with 1% SDS, 0.1 M NaHCO₃ and heated at 65 °C for 8 h to reverse the paraformaldehyde cross-linking. DNA fragments were purified with a QIAquick DNA purification kit (Qiagen) and analysed using the SYBR Green real-time PCR kit (Takara Bio Inc.). Anti-AHR (BML-SA210, Enzo Life Sciences), anti-NF- κ B p65 (D14E12) XP rabbit monoclonal antibody (8242, Cell Signaling Technology), and recombinant IgG isotype control were used as indicated in specific experiments. The following primer pairs were used: VEGFB_NFKB1: forward 5'-TCTGTGGCATAGAAACCCAAAG-3', reverse 5'-ACCCTAAGTCACTGGCTGTC-3', VEGFB_AHR1: forward 5'-ACCTTCTTCACAGGACAGCC-3' and reverse 5'-AGTCTCCGAACCTCTGTGTC-3', VEGFB_AHR2: 5'-GAGTTAACTGCAATTCCTTCACA-3' and reverse 5'-CTGGAGGGTGGTGGCTGAAG-3', VEGFB_NFK2/AHR3: forward 5'-TTCATTGGTCTCTCCCTGC-3' and reverse 5'-CAGGGGAAA GGGGACACAC-3', VEGFB_AHR3 + 4: forward 5'-GTCCCTTTTCCC CTGCAG-3' and reverse 5'-AGAGGCTCATGTGACCTAAACA-3', TGFA_AHR1: forward 5'-GCCAAGGGAGCATGAAGTAG-3' and reverse 5'-GATG CTCAAAGTTTCAGAGTTGA-3', TGFA_AHR2: forward 5'-AGGAGAGGG GTCAGTCTGAT-3' and reverse 5'-AGAGGGAAACACAAG AAGGGA-3', TGFA_AHR3: forward 5'-GACTCAGAGTGGGGCCAG-3' and reverse 5'-GAGTCGCTCAGGATCCAGTC-3'.

Human primary astrocytes and microglia. Human fetal astrocytes and microglia were isolated as previously described^{16,35} from human CNS tissue from fetuses at 17–23 weeks of gestation obtained from the Human Fetal Tissue Repository (Albert Einstein College of Medicine) and from the University of Washington Birth Defects Research Laboratory (BDRL, Project Number 5R24HD000836-51) following Canadian Institutes of Health and NIH Research-approved guidelines. Primary human astrocytes and microglia were treated with human VEGF-B (0.1 ng ml⁻¹, R&D Systems), human TGF α (0.1 ng ml⁻¹, R&D systems), or vehicle, poly(I:C) (10 mg ml⁻¹) with or without 3-indoxyl-sulfate (50 μ M, Sigma) or untreated (control) as indicated in the respective figures. After 24 h, total RNA was isolated, transcribed and subjected to qPCR.

Immunohistochemistry and astrocyte morphometry. Mice were anaesthetized (intraperitoneal injection of 100 mg ketamine and 5 mg xylazine per kg body weight) and transcardially perfused with PBS followed by 4% paraformaldehyde in PBS. CNS and other organs were post-fixed for 4–6 h at 4 °C, washed in PBS and incubated in 30% sucrose in PBS at 4 °C until fully enriched. Samples were embedded in OCT (Tissue-Tek) for frozen sectioning on a cryostat (Leica). All the stainings were performed on 30 μ m thick transversal spinal cord sections. The sections were permeabilized in blocking solution (0.5% Triton-X 100, 5% BSA, 5% normal donkey serum and 0.1% NaN₃ in PBS) for 1 h at room temperature. Primary antibodies were dissolved in blocking solution and incubated overnight at 4 °C with the following primary antibodies: Lys310-acetyl p65 (ab52175, Abcam), SOX9 (AF3075, R&D, 1:200), GFAP (RBK037, Zytomed, 1:5,000), IBA-1 (ab178846, Abcam, 1:500), DAPI (1:5,000). Conjugated secondary antibodies used were

donkey anti-rabbit Alexa Fluor 647 (Invitrogen A-31573), donkey anti-rabbit Alexa Fluor 568 (Invitrogen A-10042) and donkey anti-goat Alexa Fluor 488 (Life Technologies A11055) for 2 h at room temperature. TUNEL staining was performed using the *In situ* Cell Death Detection Kit, TMR red (Roche, 12156792910). Slices were mounted with ProLong Diamond Antifade Mountant (Life Technologies, P36961).

For immunohistochemistry with DAB, after harvesting as described above, spinal cords were cut transversally, embedded in paraffin blocks, cut in 4 μ m thick slices in a microtome, cooked 40 min in citrate buffer at 98 °C using a vapour cooker for antigen retrieval, incubated one hour using 10% normal goat serum (SouthernBiotech, 0060-01) in PBS as blocking buffer, incubated overnight at 4 °C using primary rat anti-Mac3 antibody (BD-Pharmingen Biosciences, 553322) diluted 1:200 in blocking buffer, incubated biotin-conjugated goat anti-rat antibody (SouthernBiotech, 3050-08) 45 min at room temperature diluted 1:1,000 in blocking buffer as secondary antibody and finally incubated for 45 min at room temperature with streptavidin-HRP (SouthernBiotech, 7100-05) diluted 1:1,000 in PBS. The samples were washed four times in PBS plus 0.1% Triton X-100 between each step. The DAB reaction was run for 5 min using the EnVision Flex kit (Dako). Slides were counterstained with Gill's Hematoxylin, dehydrated through an ethanol and xylol gradient and mounted with *in vitro*-Clud (R. Langenbrinck, 04-0001) using a Leica CV5030 system. All images were taken with a Keyence BZ-9000 microscope. For astrocyte reconstruction a Leica Confocal SP8 microscope was used. The Imaris 9.0.2 Software was used to reconstruct and get the morphology data.

In vivo demyelination assay. Naive mice were anaesthetized and injected stereotactically into the corpus callosum (1.4 anterior, 1.0 lateral, 2.1 mm deep) with 2 μ l of Lysolecithin (1% (w/v), Sigma), VEGF-B (500 ng, R&D Systems), or PBS. Mice were euthanized 6 days later and subjected to myelin staining as described below.

MS tissue and immunofluorescence. Brain tissue was obtained from untreated individuals with clinically diagnosed and neuropathologically confirmed MS, and healthy controls as previously described². All MS individuals and controls, or their next of kin, had given informed consent for autopsy and use of their brain tissue for research purposes. All the procedures were performed in accordance with local Institutional Review Board guidelines. MS samples were processed and immunostained as previously described². In brief, sections were thawed, fixed, washed and blocked with donkey serum 10%. Sections were then incubated overnight at 4 °C with antibodies against AHR (rabbit anti-AHR, Enzo Life Sciences), CD14 (mouse anti-human CD14, 325602, Biolegend), TGF α (ab112030, Abcam) and VEGF-B (ab185696, Abcam). After washes the samples were incubated at room temperature for 40 min with the secondary antibodies (donkey anti rabbit RRX and donkey anti mouse Alexa 488, Jackson ImmunoResearch). Imaging was performed using a Leica SP5 confocal microscope and the Leica LAS AF software. Images were processed using Adobe Photoshop CS2. For imaging analysis, all the data were acquired using the same settings, which were originally standardized on NAWM sections. The degree of co-localization of CD14 with AHR, TGF α , and VEGF-B was determined using the Volocity software from Perkin Elmer. The overlap coefficient is expressed in percentage where 100% represents the maximum degree of co-localization and 0% denotes no co-localization.

RNA-seq data processing. RNA-Seq data was analysed using DESeq2. Gene expression with 0 counts and low expression were removed before differential analysis. Low expressed were filtered out by the DESeq independent filtering, which removes genes in the lowest 40% quantile of mean normalized counts. Differential genes were selected with FDR < 0.1.

Heat maps. Heat maps were generated with the Gene-E program, and the z-scores were calculated for each gene row using the mean expression of biological replicates.

tSNE plots. tSNE plot was created using R and the Rtsne package, with parameters perplexity = 1, max iterations = 3,000. The mean average of replicates and the top 500 ranked genes for TDD + Trp and TDD in control mice ($n = 3$) was taken and the final plot was generated using ggplot.

Ingenuity pathway analysis. To determine significant pathways, differentially expressed genes that passed FDR < 0.1 were uploaded and analysed using Ingenuity Pathway Analysis (IPA) tool. *P* values of canonical signalling pathways were calculated using Fischer's exact test. The NF- κ B network diagram was generated using IPA.

Network diagram. Network diagram for protein–protein interactions was visualized with NetworkAnalyst, (<http://www.networkanalyst.ca>), using the STRING Interactome database (confidence score cutoff = 900). Minimum network displaying interacting mediators and molecules were colour-coded based on associated pathways for VEGF-B and TGF α .

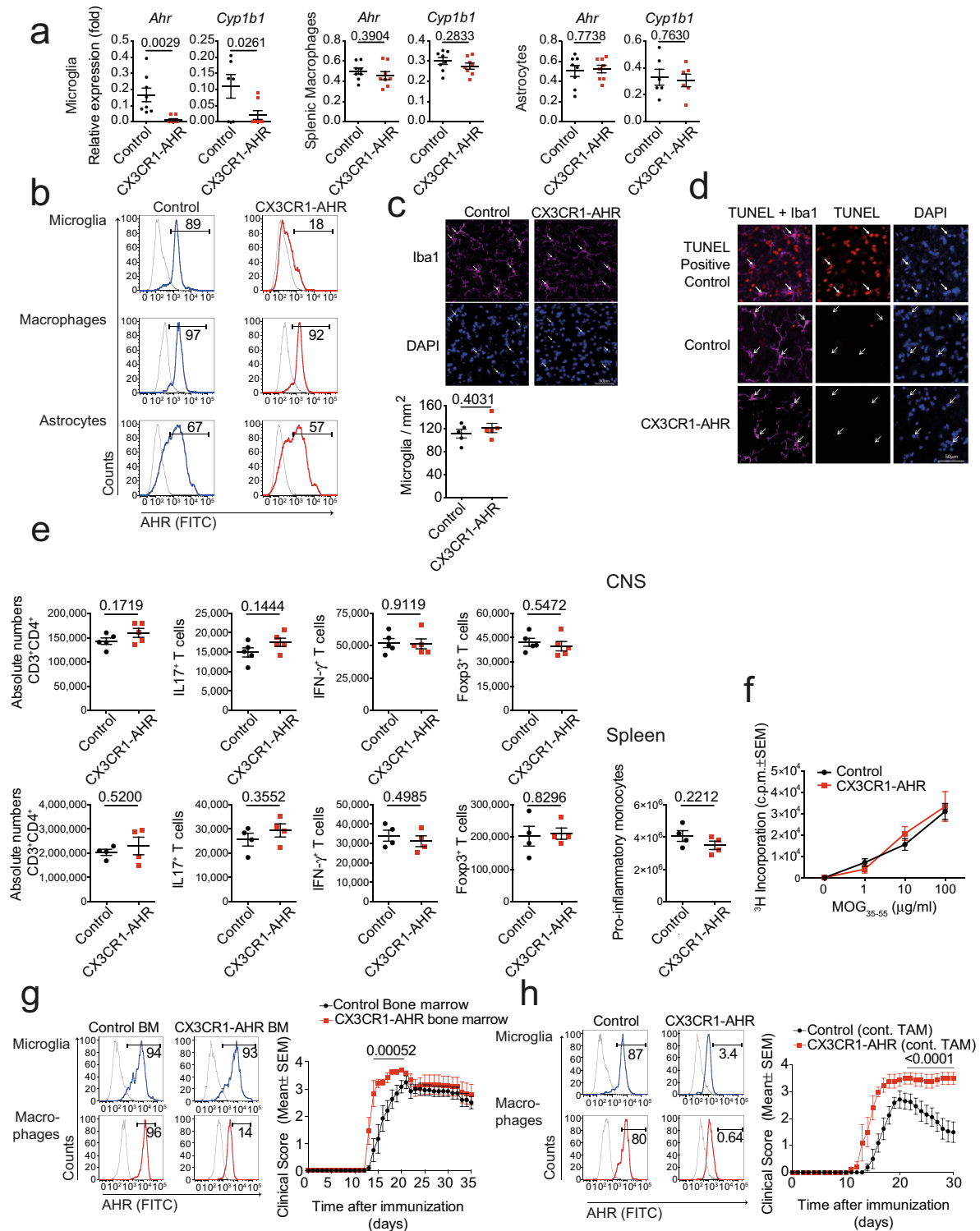
Statistical analysis. Statistical analyses were performed with Prism software (GraphPad), using the statistical tests indicated in the individual figure legends. No samples were excluded. The investigators were blinded as to the treatment

of mice in individual experiments. $P < 0.05$ was considered significant. All error bars represent s.e.m. or s.d. as noted in the individual figure legends. Unless otherwise stated, three independent experiments were used for all assays, and displayed figures are representative. The experiments were not randomized, and no statistical methods were used to predetermine sample size.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. RNA-seq data have been uploaded and are accessible under the access code <https://figshare.com/s/c109f251b149b7a843b3>.

31. Mildner, A. et al. Distinct and non-redundant roles of microglia and myeloid subsets in mouse models of Alzheimer's disease. *J. Neurosci.* **31**, 11159–11171 (2011).
32. Mildner, A. et al. Microglia in the adult brain arise from Ly-6C^{hi}CCR2⁺ monocytes only under defined host conditions. *Nat. Neurosci.* **10**, 1544–1553 (2007).
33. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protocols* **7**, 562–578 (2012).
34. Yan, Y. et al. CNS-specific therapy for ongoing EAE by silencing IL-17 pathway in astrocytes. *Mol. Ther.* **20**, 1338–1348 (2012).
35. Jack, C. S. et al. TLR signaling tailors innate immune responses in human microglia and astrocytes. *J. Immunol.* **175**, 4320–4330 (2005).



Extended Data Fig. 1 | See next page for caption.

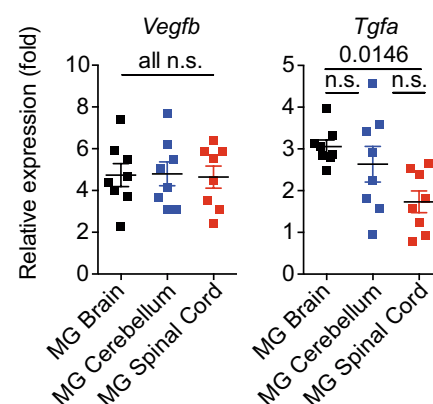
Extended Data Fig. 1 | Contribution of AHR in CNS resident and infiltrating immune cells during EAE. **a**, qPCR of indicated genes from microglia, splenic macrophages, and astrocytes from control and CX3CR1-AHR mice on day 28 after EAE induction. $n = 8$ independent samples per group. **b**, Flow cytometry analysis of AHR expression in microglia, monocytes and astrocytes from Control and CX3CR1-AHR mice 21 days after EAE induction. Thin line depicts isotype control, thick line AHR staining, and numbers indicate percentage of AHR positive cells. Representative of stainings of $n = 3$ mice per group. **c**, Spinal cord samples from naive Control and CX3CR1-AHR mice were stained for Iba-1 and DAPI and Iba-1⁺ microglia/mm² were determined. $n = 5$ mice per group. n.s., not significant. **d**, TUNEL staining in Iba-1⁺ microglia in spinal cord sections of control and CX3CR1-AHR mice as in **c**. For the positive control, slides were cooked at 98 °C in citrate buffer during 60 min using a vapour cooker. Solid arrows show TUNEL positive microglia. Representative of $n = 5$ independent experiments. **e**, Number of CNS-infiltrating (top) and splenic T cells (bottom), and splenic pro-inflammatory monocytes (bottom) as determined by flow cytometry. $n = 5$ samples per group for CNS, $n = 4$ samples per group for spleen. **f**, Proliferation assay from splenocytes isolated on day 28 of the experiment. $n = 4$ biologically independent samples per group,

representative of two independent experiments. **g**, Bone marrow chimaera were generated using wild-type mice irradiated as recipients, reconstituted with control or CX3CR1-AHR bone marrow. Recipients of bone marrow were then rested for 3 weeks and thereafter treated with weekly tamoxifen gavages (4 mg) for another 3 weeks; after a total of 6 weeks, EAE was induced and tamoxifen administration continued weekly during EAE. Left, flow cytometry analysis of AHR expression in microglia and monocytes 21 days after EAE induction. Thin line depicts isotype control, thick line denotes AHR staining, and numbers indicate the percentage of AHR-positive cells. Representative of stainings of $n = 3$ independent mice per group. Right, EAE clinical course in bone marrow chimaera mice. $n = 4$ mice per group. **h**, Control and CX3CR1-AHR mice were treated with oral tamoxifen weekly starting from 5 weeks of age. EAE was induced at 8 weeks under continuation of weekly tamoxifen administration. Left, intracellular FACS staining for AHR in microglia and monocytes from at day 21 of EAE. Representative of stainings of $n = 3$ independent mice per group. Right, clinical course of control and CX3CR1-AHR bone marrow chimaera mice. Data in **a**, **c**, **e–h** are mean \pm s.e.m. of $n = 4$ mice per group. *P* values were determined by two-sided Student's *t*-test (**a**, **c**, **e**) or two-way ANOVA (**g**, **h**).

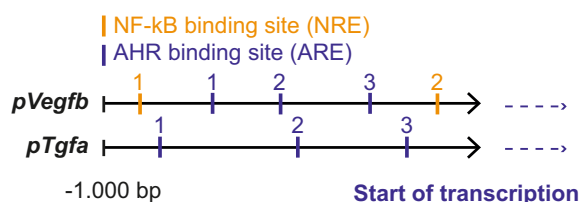
a

Ingenuity Canonical Pathways	P-value
Axonal Guidance Signaling	0.000003
NF- κ B Signaling	0.000282
ErbB Signaling	0.001072
Production of Nitric Oxide and Reactive Oxygen Species	0.004898
VEGF Family Ligand-Receptor Interactions	0.008128
VEGF Signaling	0.018621
Leukocyte Extravasation Signaling	0.022387
Communication between Innate and Adaptive Immune Cells	0.028840

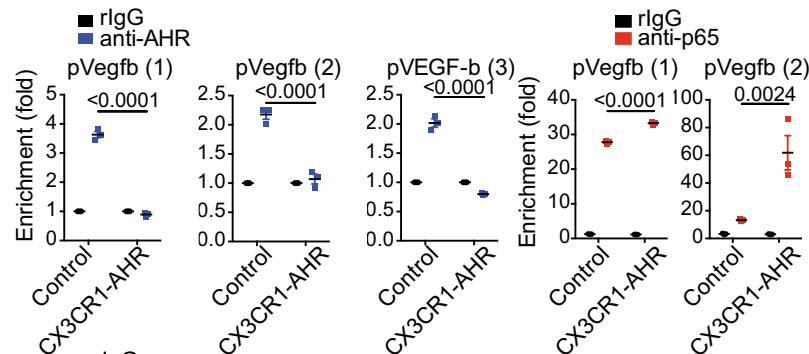
b



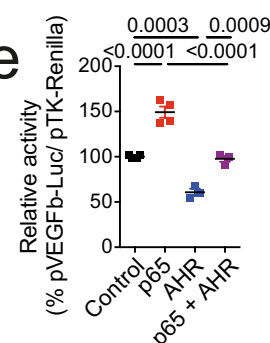
c



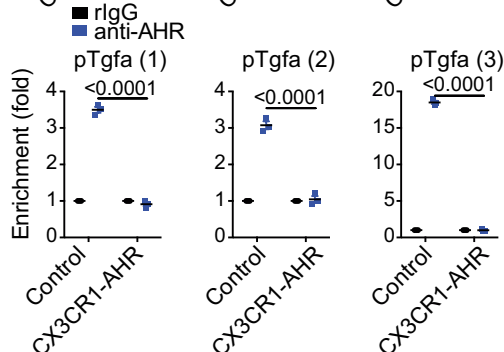
d



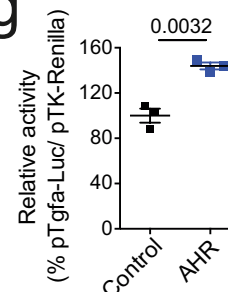
e



f

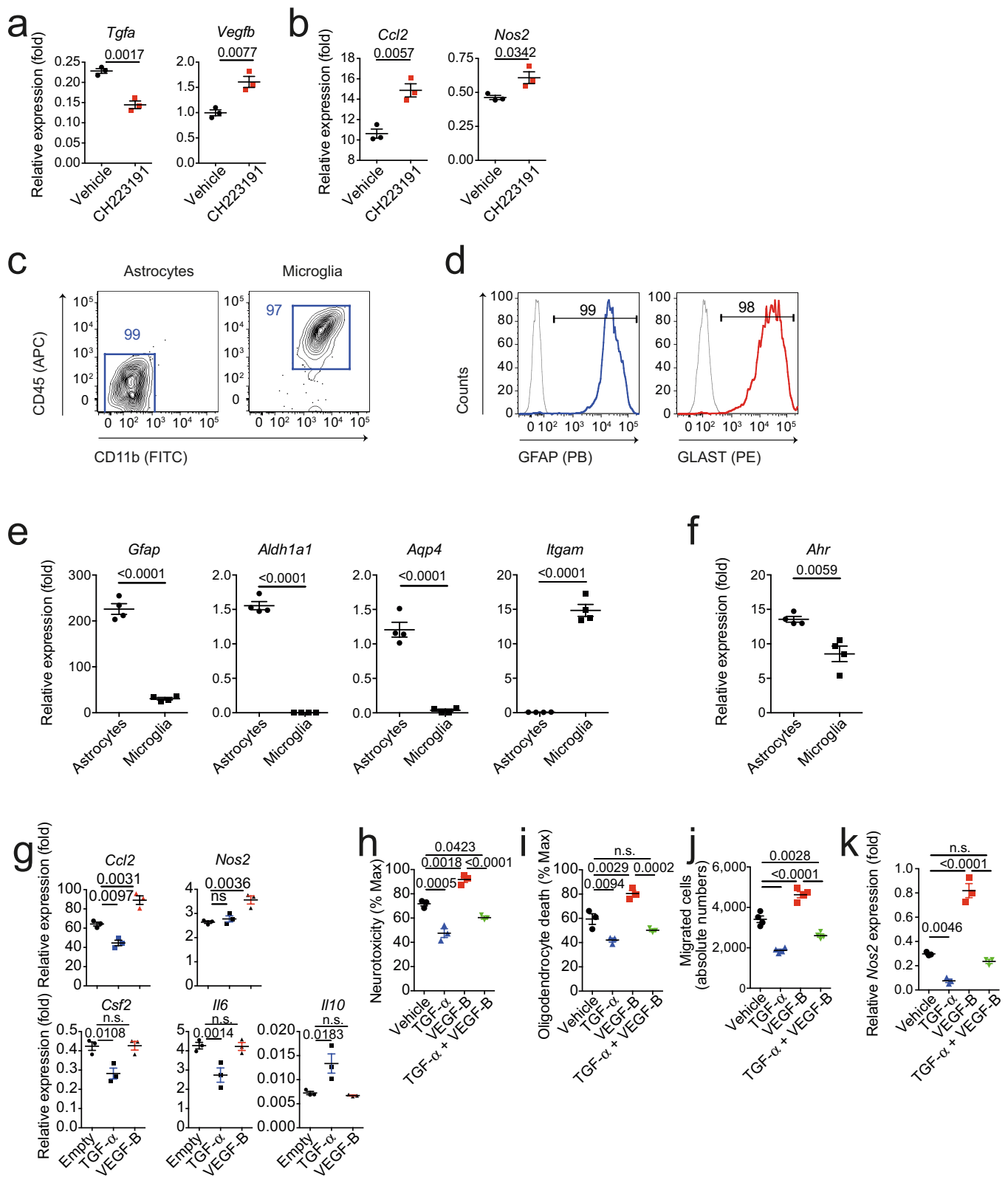


g



Extended Data Fig. 2 | Topical and molecular regulation of TGF α and VEGF-B. **a**, Ingenuity pathway analysis of differentially regulated pathways in astrocytes from $n = 3$ control versus CX3CR1-AHR mice per group during EAE. **b**, *Tgfa* and *Vegfb* expression determined by qPCR in microglia from brain, cerebellum and spinal cord 21 days after EAE induction (left). $n = 8$ mice per group. **c**, Predicted NF- κ B and AHR responsive sites (NREs and XREs, respectively) in *Vegfb* and *Tgfa* promoters. **d**, Microglia were isolated by FACS sorting from control and CX3CR1-AHR mice during EAE. Ex vivo ChIP assay of NF- κ B p65 or AHR binding to predicted binding sites in the *Vegfb* promoter. $n = 3$ mice per group. Data are representative of two independent experiments. **e**, Reporter assay using a construct in which the *Vegfb* promoter controls luciferase expression (pVegfb-Luc). Luciferase activity was measured in

HEK293 cells 24 h after transfection with pVegfb-Luc, pTK-Renilla, and plasmids expressing AHR or NF- κ B p65. Data are representative of two independent experiments with four biological replicates. **f**, Ex vivo ChIP assay as in **d** for AHR binding to the *Tgfa* promoter. $n = 3$ mice per group. Representative of two independent experiments. **g**, Reporter assays using a construct in which the *Tgfa* promoter controls luciferase expression (pTgfa-Luc). Luciferase activity was measured in HEK293 cells 24 h after transfection with pTgfa-Luc, pTK-Renilla, and plasmids expressing AHR or control. Data are representative of two independent experiments with three biological replicates. Data in **a**, **d**–**g** are mean \pm s.e.m. P values were determined by one-way ANOVA followed by Tukey's post-hoc test (**b**, **d**, **e**, **f**) or two-sided Student's t -test (**g**).

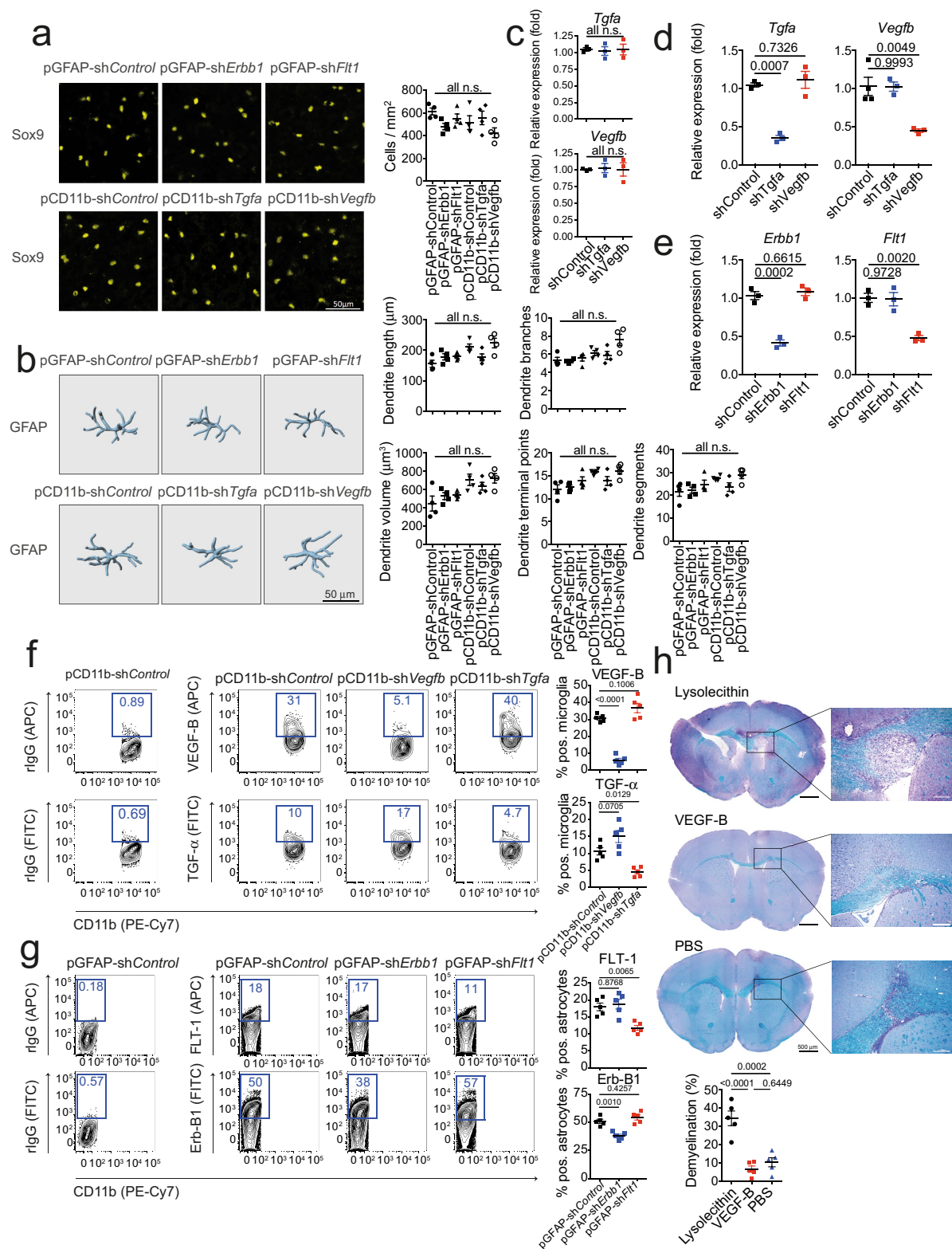


Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | TGF α and VEGF-B are regulated by AHR in highly purified astrocytes and microglia. **a, b**, Mouse microglia were activated with lipopolysaccharide (LPS) in the presence or absence of the AHR inhibitor CH223191. After 24 h, activation medium was removed and substituted with fresh medium after extensive washes. Then 48 h later, microglia conditioned medium (MCM) was collected and applied to cultures of primary astrocytes. **a**, Gene expression in microglia 24 h after activation in the presence or absence of CH223191. **b**, Gene expression in astrocytes after 24 h exposure to MCM. Data are representative of two independent experiments with three biological replicates. **c**, Representative FACS stainings for CD11b and CD45 in primary astrocyte and microglia cultures. Numbers indicate percentages in respective gate. Data are representative of three independent experiments. **d**, Representative FACS stainings for GFAP and GLAST in astrocyte cultures as in **b**. Data are representative of three independent experiments. **e, f**, qPCR analysis of mRNA expression in astrocyte and microglia cultures. $n = 4$ independent cultures. Data are representative of two independent experiments with four biological replicates. **g**, Effect of TGF α

and VEGF-B on gene expression in primary astrocytes activated with TNF and IL-1 β , determined by pPCR after 24 h. Data are representative of three independent experiments with three biological replicates.

h, i, Primary mouse astrocytes were activated with TNF and IL-1 β and treated with TGF α or VEGF-B. After 24 h later, culture medium was substituted by fresh medium after extensive washes. Then 48 h later, ACM was added to mouse neurons (**h**) and oligodendrocytes (**i**) in culture, and cytotoxicity was determined by quantifying lactate dehydrogenase (LDH) release after 24 h. $n = 3$ biological replicates. Data are representative of two independent experiments. **j**, CD11b⁺Ly6C^{hi} monocyte migration assay performed using ACM from astrocytes activated in the presence of TGF α or VEGF-B. $n = 4$ biological replicates. Data are representative of two independent experiments. **k**, qPCR analysis of *Nos2* expression in microglia co-cultured with astrocytes activated in the presence of TGF α or VEGF-B. $n = 3$ biological replicates. Data are representative of two independent experiments. Data in **b, e–k** are mean \pm s.e.m. *P* values were determined by two-sided Student's *t*-test (**b, e, f**) or one-way ANOVA followed by Tukey's post-hoc test (**g–k**).

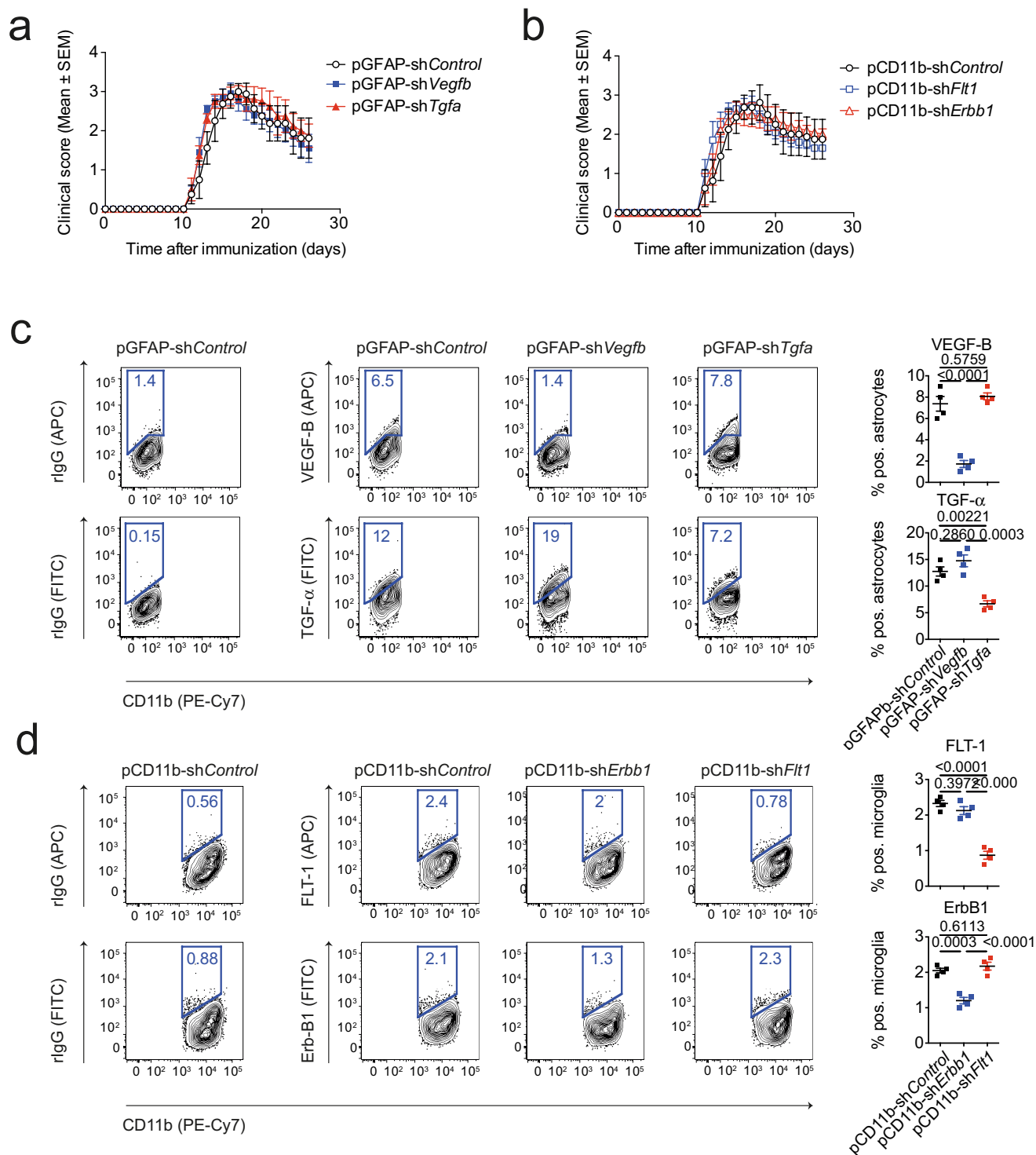


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Phenotypical and functional effects of

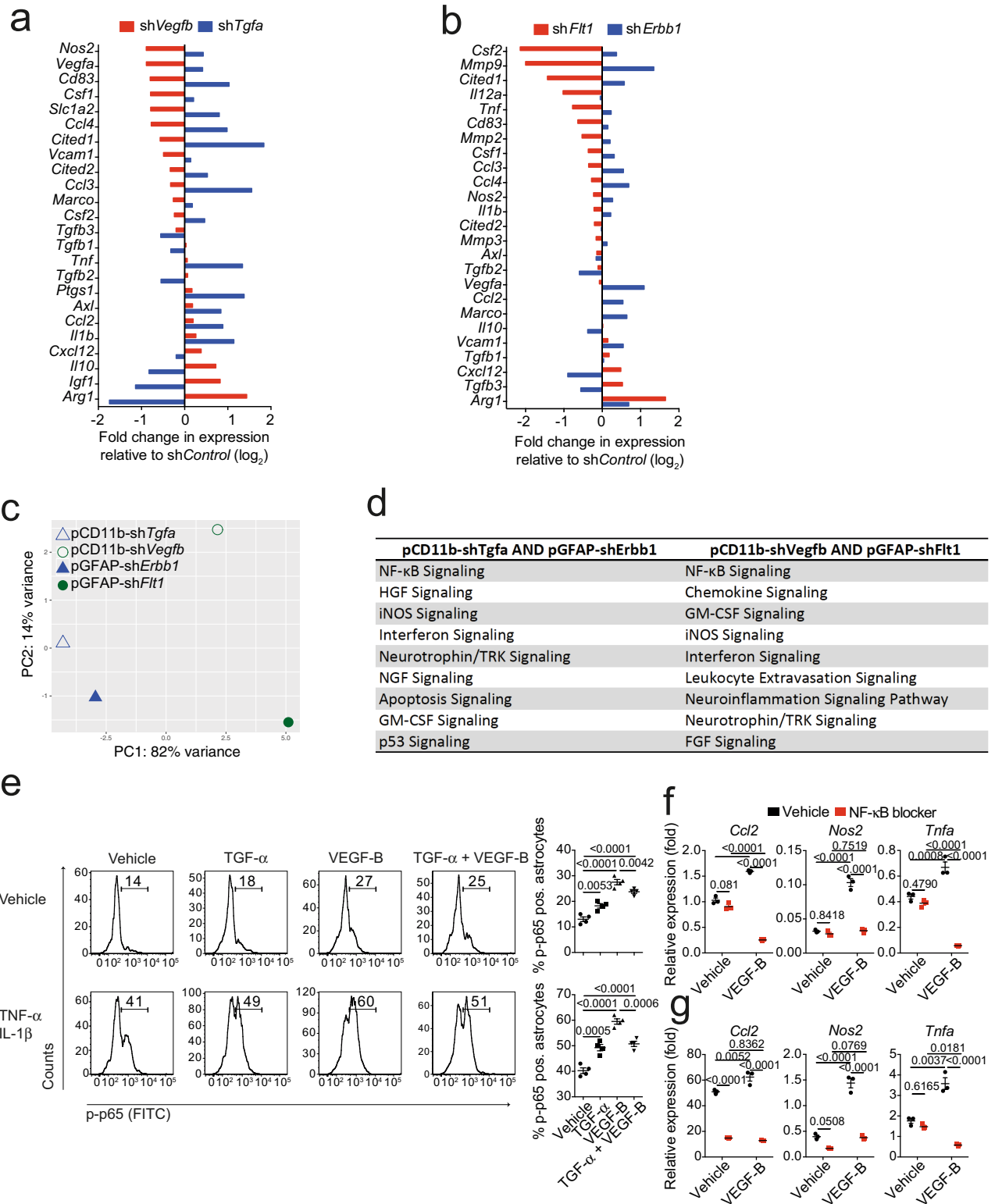
knockdown of microglial TGF α and VEGF-B. **a**, Quantification of astrocyte numbers in spinal cord sections of knockdown mice. SOX9-positive astrocytes per mm² were quantified in spinal cord sections of four mice per group. **b**, IMARIS reconstruction of GFAP⁺ astrocytes in spinal cord sections as in **a**, and quantification of dendrite length, branches, volume, terminal points, and segments of $n = 4$ mice per group. **c**, **d**, qPCR analysis of *Tgfa* and *Vegfb* expression in sorted CNS-infiltrating inflammatory monocytes (**c**) and microglia (**d**) from mice injected with pCD11b-shControl, pCD11b-shTgfa, and pCD11b-shVegfb 7 days after EAE induction. Representative of two independent experiments with three biological replicates. **e**, qPCR analysis of *Erbb1* and *Flt1* expression in mice injected with pGFAP-shControl, pGFAP-shErbb1, and pCD11b-shFlt1 7 days after EAE induction. Representative of two independent experiments with three biological replicates. **f**, Left, flow cytometry analysis of VEGF-B

and TGF α expression in microglia from mice injected with pCD11b-shControl, pCD11b-shTgfa, and pCD11b-shVegfb 7 days after EAE induction. Right, quantification of VEGF-B- and TGF α -positive microglia in $n = 5$ mice per group. Representative of two independent experiments with five biological replicates. **g**, Left, flow cytometry analysis of FLT-1 and ErbB1 expression in astrocytes from mice injected with pGFAP-shControl, pGFAP-shErbb1, and pCD11b-shFlt1 7 days after EAE induction. Right, quantification of FLT-1 and ErbB1-positive microglia in $n = 5$ mice per group. Representative of two independent experiments with five biological replicates. **h**, Naive mice were injected with lysolecithin, VEGF-B, or PBS into the corpus callosum by stereotaxic injection and 6 days later, brains were analysed by myelin staining. Representative of two independent experiments with five biological replicates. Data are mean \pm s.e.m. *P* values were determined one-way ANOVA followed by Tukey's post-hoc test (**a–h**).



Extended Data Fig. 5 | Directionality of TGF α and VEGF-B signalling during EAE. **a, b**, EAE development in wild-type mice injected with pGFAP-shControl, pGFAP-shErbB1 and pCD11b-shFlt1 (**a**), or pCD11b-shControl, pCD11b-shTgfa and pCD11b-shVegfb (**b**). Clinical course. $n = 5$ mice per group. Representative of two independent experiments with $n = 5$ mice per group. **c**, Left, flow cytometry analysis of TGF α and VEGF-B expression in astrocytes as in **a**. Right, quantification of cytokine-positive astrocytes. Data are mean \pm s.e.m.

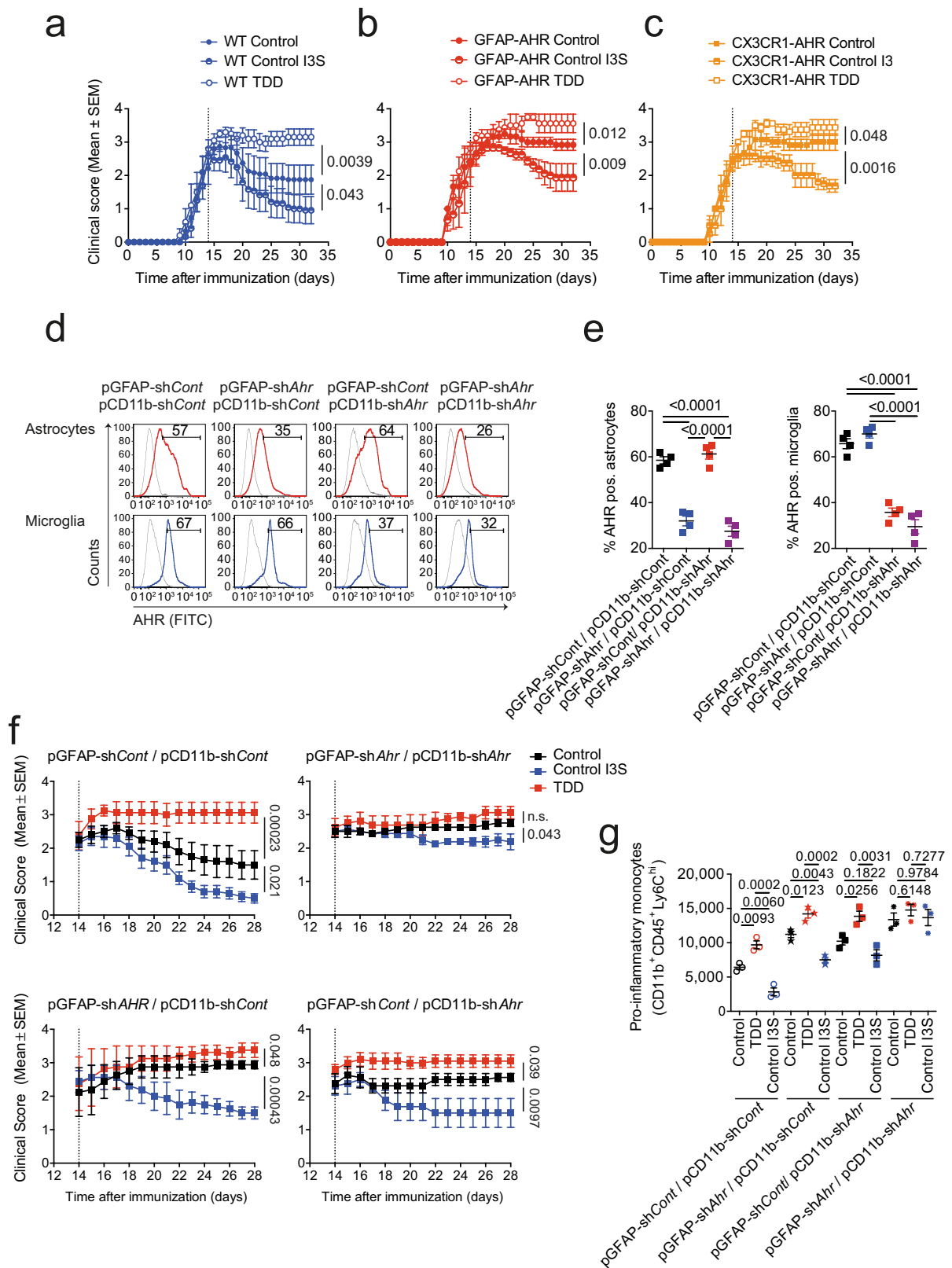
and P values were determined by one way ANOVA followed by Tukey's post-hoc test. Representative of two independent experiments with four biological replicates. **d**, Left, flow cytometry analysis of FLT-1 and ErbB1 expression in microglia as in **b**. Right, quantification of cell-surface receptor expression of microglia. Data are mean \pm s.e.m. and P values were determined by one way ANOVA followed by Tukey's post-hoc test. Representative of two independent experiments with four biological replicates.



Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Regulation and transcriptional effects of TGF α and VEGF-B during EAE. **a, b**, NanoString analysis of mRNA expression in astrocytes from EAE mice injected with pCD11b-shVegfb or pCD11b-shTgfa (**a**) and pGFAP-shFlt1 or pGFAP-shErbb1 (**b**; see also Fig. 2k, l). Fold change in relative expression relative to control as determined by $\log_2(\text{shKD/shControl})$. shKD, shRNA knockdown. Representative of two independent experiments with pooled RNA isolated from $n = 3$ mice per group. **c**, Principal component analysis of gene expression in astrocytes isolated as in **a** and **b**. Representative of two independent experiments with pooled RNA isolated from $n = 3$ mice per group. **d**, Ingenuity pathway analysis of significantly regulated pathways from astrocytes as in **a** and **b**. Representative of two independent experiments with pooled RNA isolated from $n = 3$ mice per group. **e**, Left, representative flow cytometry plots depicting NF- κ B p65 phosphorylation in wild-type astrocytes stimulated for 15 min with vehicle (top) or TNF or IL-1 β (bottom) in the presence of TGF α , VEGF-B, or their combination.

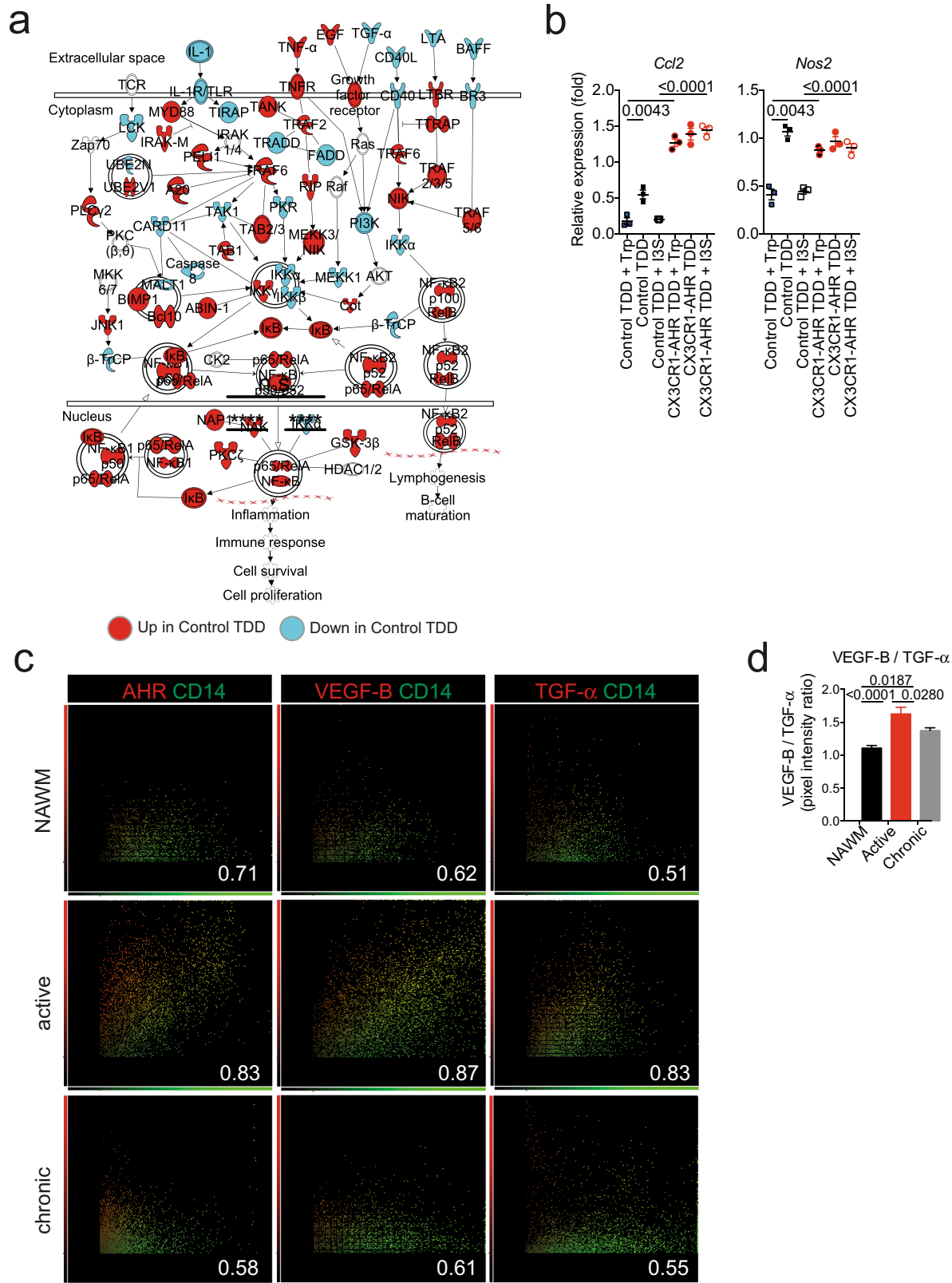
Numbers indicate percentage of FITC $^{+}$ cells. Bar graphs depict quantification of FITC $^{+}$ cells. Data are mean \pm s.e.m. and P values were determined by one way ANOVA followed by Tukey's post-hoc test. Representative of two independent experiments with four biological replicates. **f**, Primary mouse astrocytes were exposed to VEGF-B or vehicle and pharmacological blocker of NF- κ B activation. RNA was obtained after 18 h and subjected to qPCR analyses for the indicated genes. Data are mean \pm s.e.m. and P values were determined by one way ANOVA followed by Tukey's post-hoc test. Representative of two independent experiments with three biological replicates. **g**, Primary mouse astrocytes were activated with TNF or IL-1 β in the presence of VEGF-B or vehicle, and a pharmacological blocker of NF- κ B activation. RNA was obtained after 18 h and subjected to qPCR analyses for the indicated genes. Data are mean \pm s.e.m. and P values were determined by one way ANOVA followed by Tukey's post-hoc test. Representative of two independent experiments with $n = 3$ biological replicates.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Role of AHR in astrocytes and microglia during EAE. **a–c**, EAE was induced in control (WT), *Gfap^{cre}Ahr^{fl/fl}* (GFAP-AHR), or CX3CR1-AHR EAE mice. Starting from day 7, mice were injected daily intraperitoneally with indoxyl-3-sulfate (I3S), given a tryptophan-depleted diet (TDD), or kept on a control diet. Clinical course of EAE mice under treatment conditions as indicated. Representative of two independent experiments with $n = 4$ mice per group. **d–f**, EAE was induced in wild-type mice, which were treated with lentiviruses to knockdown AHR in astrocytes (pGFAP-shAhr) or microglia (pCD11b-shAhr). A noncoding RNA was used as a control. Flow cytometry quantification of AHR expression in astrocytes and microglia by FACS. **d**, Representative histograms of $n = 4$ mice per group. Numbers indicate percentage of AHR-positive cells; thin lines denote isotype control, thick lines denote

AHR staining. **e**, Quantification of AHR-positive astrocytes and microglia as in **d**. Representative of two independent experiments with four biological replicates. **f**, EAE mice with knock down of AHR in astrocytes, microglia, or both as in **d** were subjected to daily I3S injections, TDD, or control diet conditions starting on day 14 after disease induction. Clinical course of $n = 4$ mice per group. Representative of two independent experiments with $n = 4$ mice per group. **g**, Quantification of CNS-infiltrating pro-inflammatory monocytes as determined by FACS at day 28 of EAE. Representative of two independent experiments with three biological replicates. Data are mean \pm s.e.m. *P* values were determined by two-way ANOVA (**a**, **f**), or one way ANOVA followed by Tukey's post-hoc test (**e**, **g**).



Extended Data Fig. 8 | Dietary factors influence mouse and human TGF α and VEGF-B expression. **a**, Ingenuity pathway analysis of NF- κ B signalling comparing a TDD to a TDD plus Trp diet in control animals. Colours code for up- and downregulation of individual members in red (up) and blue (down). Normalized reads of two independent samples per group. **b**, mRNA expression determined by qPCR in from EAE mice as in Fig. 3a. Data are representative of two independent experiments with three replicates. **c**, Quantification of co-expression of AHR and CD14, VEGF-B

and CD14, TGF α and CD14 in immunofluorescence stainings of human white matter brain tissue of NAWM, active, or chronic MS lesions for AHR (left), VEGF-B (middle), or TGF α (right), CD14 (green), and DAPI (blue). Data are representative of $n = 12$ fields from three distinct MS brains. **d**, Ratio of VEGF-B to TGF α intensities. Data are the ratio of mean values from Fig. 4e + s.e.m. of $n = 25$ fields. Data in **b** and **d** are mean \pm s.e.m. P values derived by one-way ANOVA followed by Tukey's post-hoc test (**b**, **d**).

Mechanism of phosphoribosyl-ubiquitination mediated by a single *Legionella* effector

Anil Akturk^{1,4}, David J. Wasilko^{1,4}, Xiaochun Wu¹, Yao Liu², Yong Zhang², Jiazhang Qiu², Zhao-Qing Luo², Katherine H. Reiter³, Peter S. Brzovic³, Rachel E. Klevit³ & Yuxin Mao^{1*}

Ubiquitination is a post-translational modification that regulates many cellular processes in eukaryotes^{1–4}. The conventional ubiquitination cascade culminates in a covalent linkage between the C terminus of ubiquitin (Ub) and a target protein, usually on a lysine side chain^{1,5}. Recent studies of the *Legionella pneumophila* SidE family of effector proteins revealed a ubiquitination method in which a phosphoribosyl ubiquitin (PR-Ub) is conjugated to a serine residue on substrates via a phosphodiester bond^{6–8}. Here we present the crystal structure of a fragment of the SidE family member SdeA that retains ubiquitination activity, and determine the mechanism of this unique post-translational modification. The structure reveals that the catalytic module contains two distinct functional units: a phosphodiesterase domain and a mono-ADP-ribosyltransferase domain. Biochemical analysis shows that the mono-ADP-ribosyltransferase domain-mediated conversion of Ub to ADP-ribosylated Ub (ADPR-Ub) and the phosphodiesterase domain-mediated ligation of PR-Ub to substrates are two independent activities of SdeA. Furthermore, we present two crystal structures of a homologous phosphodiesterase domain from the SidE family member SdeD⁹ in complexes with Ub and ADPR-Ub. The structures suggest a mechanism for how SdeA processes ADPR-Ub to PR-Ub and AMP, and conjugates PR-Ub to a serine residue in substrates. Our study establishes the molecular mechanism of phosphoribosyl-linked ubiquitination and will enable future studies of this unusual type of ubiquitination in eukaryotes.

A variety of microbial pathogens exploit the eukaryotic ubiquitination pathway during their respective infections^{10,11}. The intracellular pathogen *L. pneumophila* injects more than 300 effectors into host cells during its infection, including at least ten proteins that are involved in ubiquitin manipulation¹². These effectors include HECT-like^{13,14} and F- or U-box-containing Ub ligases^{15–18} as well as novel Ub ligases of the SidE family, such as SdeA, that act independently of canonical E1 and E2 enzymes^{6–8}. SdeA first uses its mono-ADP-ribosyltransferase (mART) activity to catalyse the transfer of ADP-ribose from NAD⁺ to the side chain of R42 on Ub to generate ADPR-Ub. Subsequently, SdeA uses its phosphodiesterase (PDE) activity to catalyse the conjugation of ADPR-Ub to a serine residue on substrates to generate a protein–PR-Ub product. Alternatively, in the absence of substrates, the SdeA PDE domain will catalyse the hydrolysis of ADPR-Ub to generate PR-Ub and AMP (Fig. 1a, Extended Data Fig. 1). The molecular mechanism of this unique ubiquitination pathway is still unknown.

To determine the mechanism of phosphoribosyl-linked ubiquitination, we determined the crystal structure of a portion of SdeA (amino acids 211–910, hereafter called SdeA-core; Extended Data Table 1). The structure is composed of two distinct domains, the PDE and mART domains (Fig. 1b, c). A calculation of the surface electrostatic potential revealed no notably charged areas on the surface of SdeA other than a deep and highly positively charged groove on the PDE domain (Fig. 1d, e). Analogous to other PDEs¹⁹, the active site is likely to be harboured in this deep groove (Extended Data Fig. 2a–c). Indeed, a sequence

alignment of PDE domains showed that most of the conserved residues reside in this groove, consistent with their forming the PDE active site (Extended Data Figs. 2d, 3). The mART domain is composed of two lobes, an N-terminal α -helical lobe (amino acids 592–758) and a main lobe (amino acids 759–911). The main lobe contains a β -sandwich core and harbours the three catalytic motifs: the (F/Y)-(R/H), STS and EXE motifs (Extended Data Figs. 4a–f, 5) that are conserved in other mART proteins, such as the *Pseudomonas syringae* effector HopU1 and the *Clostridium perfringens* iota-toxin^{20–22}. A structural comparison of the α -helical lobe with its counterparts in other mARTs revealed that although the total number and the length of α -helices are variable, three α -helices form a structural core that is conserved in most mART proteins (Extended Data Fig. 4g–i). Although it packs in close contact with the main lobe in other mARTs, the α -helical lobe is extended away from the main lobe in our SdeA-core crystal structure (Extended Data Fig. 6a, b). The extended conformation observed in our crystal structure is consistent with the conformation in solution as detected by small-angle X-ray scattering (SAXS) and does not change in the presence of NAD⁺ (Extended Data Fig. 6c–f). However, the α -helical lobe adopts a closed conformation and mediates contact with NAD⁺ in a structure of iota-toxin²¹. Moreover, the α -helical lobe is enriched with highly conserved residues (including N723, Q727 and R729) that form a cluster on its surface, as revealed by an analysis of surface residue conservation using the ConSurf server²³ (Extended Data Figs. 5, 7a). Thus, we hypothesized that the α -helical lobe may have a similar role in SdeA catalysis. Indeed, an α -helical lobe deletion in SdeA (SdeA- $\Delta\alpha$ -lobe), as well as N723A, Q727A or R729A point mutations in the α -helical lobe completely abrogated ADP-ribosylation activity (Extended Data Fig. 7b, c). A mutation in a residue that is not conserved but is close to the conserved surface patch (F719A), yielded a substantial impairment of activity, whereas mutation of a conserved residue that is away from the patch (D622A) resulted in an activity level comparable to wild-type SdeA. Taken together, our data show that the α -helical lobe is crucial for ADP-ribosylation of Ub, and that a surface patch composed of highly conserved residues may mediate the binding of NAD⁺ during catalysis. These observations further suggest that the closed conformation of the α -helical lobe is required for the mART activity of SdeA. An accompanying paper describing the crystal structure of a longer construct of SdeA in complex with both NAD⁺ and Ub reports that the α -helical lobe is indeed observed in a closed conformation²⁴.

The main lobe of the mART domain is packed against the PDE domain in the SdeA structure. The two catalytic sites face in opposite directions and are separated by a distance of over 55 Å (Fig. 1b), which raises the question of how the activities of the two domains are coordinated. To address this question, we performed assays with SdeA fragments that retain only mART or PDE activity (Fig. 2a). Similar to wild-type SdeA-core, reactions that contain both SdeA-PDE and SdeA-mART efficiently generate PR-Ub and ubiquitinate the substrate RAB33B (Fig. 2b, c). SdeA-core carrying a mutation (H277A) in the PDE active site retained the ability to generate ADPR-Ub but failed

¹Weill Institute for Cell and Molecular Biology and Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA. ²Purdue Institute of Immunology, Inflammation and Infectious Disease and Department of Biological Sciences, Purdue University, West Lafayette, IN, USA. ³Department of Biochemistry, University of Washington, Seattle, WA, USA. ⁴These authors contributed equally: Anil Akturk, David J. Wasilko. *e-mail: ym253@cornell.edu

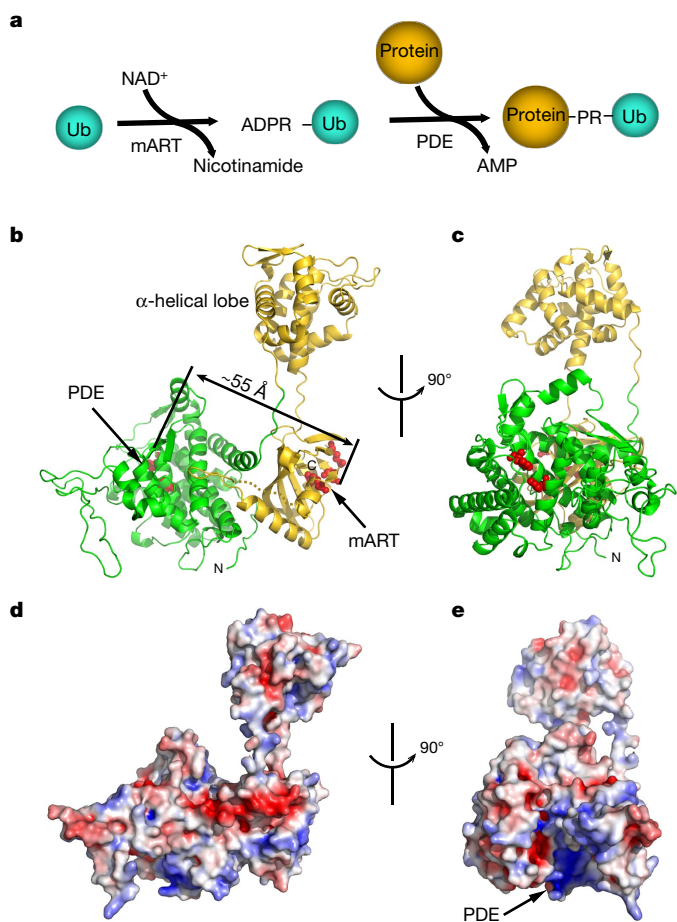


Fig. 1 | Overall structure of SdeA. **a**, Schematic of the phosphoribosyl-ubiquitination reaction. **b**, Overall structure of SdeA-core in ribbon representation. This portion of SdeA has two distinct domains: the PDE (green) and mART (gold) domains. The active site residues of both the mART and PDE domains are shown as red spheres. The linear distance between these two active sites is approximately 55 Å. **c**, An orthogonal view of **b**. **d**, Molecular surface model of SdeA. The surface is coloured on the basis of electrostatic potential with positively charged regions in blue and negatively charged surfaces in red. The orientation of the molecule is the same as shown in **b**. **e**, An orthogonal view of **d**.

to process ADPR-Ub to PR-Ub or to ubiquitinate RAB33B. However, the presence of both SdeA-core^{H277A} and SdeA-PDE successfully catalysed both the production of PR-Ub and the ubiquitination of RAB33B. Moreover, SdeA-PDE alone can catalyse phosphoribosyl-linked ubiquitination of RAB33B when purified ADPR-Ub is supplied (Fig. 2d). The independence of the two activities was further validated by SdeA-mediated RAB33B ubiquitination when the PDE and mART domains were co-expressed in cells (Fig. 2e). These results suggest that ADP-ribosylation of Ub and phosphoribosyl-linked ubiquitination of serine are mechanistically and spatially independent activities performed by a single protein.

Despite sharing 23% sequence similarity with a well-characterized cyclic di-3',5'-GMP phosphodiesterase in *Pseudomonas aeruginosa* PA4781²⁵, the PDE domain of SdeA uses ADPR-Ub as its substrate and catalyses the unprecedented phosphoribosyl-linked ubiquitination of serine. To understand how ADPR-Ub is recognized by the SdeA PDE domain, we assessed the interaction of Ub and several homologous PDE domains from the *Legionella* SidE-effector family using ¹H-¹⁵N HSQC TROSY (heteronuclear single quantum coherence, transverse relaxation-optimized spectroscopy) NMR titration experiments (Extended Data Fig. 8a–c). The SdeA PDE domain showed no detectable interaction with Ub in solution, whereas the PDE domain of another SidE family member, SdeD, exhibited a direct and specific interaction with

Ub as evidenced by NMR-peak perturbations. We then successfully determined the structures of SdeD, both on its own and in complex with Ub (Extended Data Fig. 8d–f). Notably, two Ub molecules are in contact with a single PDE domain in the crystal. One Ub (Ub2) binds on the opposite side to the catalytic groove, making the physiological significance of this binding mode unclear (Extended Data Fig. 8g). The other Ub (Ub1) binds to a flat surface at the opening of the catalytic groove (Fig. 3a). Similar to the Ub surface area mapped by NMR titration experiments in solution (Extended Data Fig. 8c), three regions of Ub1 contact the PDE domain: the loop region around residue T9, the C terminus and a region that includes R42 (Fig. 3a). At the T9 loop region, in addition to the hydrophobic interactions mainly contributed by L8, residue K6 of Ub1 forms electrostatic interactions with E251 on SdeD (Fig. 3b). At the C terminus of Ub1, in addition to hydrophobic interactions mediated by L73, R72 of Ub1 forms salt bridges with E242 on SdeD (Fig. 3c). Notably, the R42 side chain of Ub1 extends into the catalytic groove and forms hydrogen bonds and electrostatic interactions with the conserved residues Q52 and E126 at the PDE catalytic site (Fig. 3d). To test whether the PDE domain of SdeA interacts with Ub in a manner that is similar to SdeD, we modelled Ub binding by the PDE domain of SdeA on the basis of the SdeD–Ub1 complex (Fig. 3e). The model predicts that E465 and E454 in SdeA would have analogous roles in Ub binding to E251 and E242 in SdeD, respectively (Fig. 3a, e). Consistent with this prediction, PDE activity was substantially impaired in SdeA E465A and E454A mutants as evidenced by the marked reduction of both the Pro-Q staining signal and ubiquitination of RAB33B (Fig. 3f, g). In addition, a V414Y mutant designed to sterically block the access of ADPR-Ub to the catalytic site also largely impaired the PDE activity (Fig. 3e–g). All three SdeA mutants were able to cause a band shift of Ub on native gels (Fig. 3e, top) indicating that the mART activity of these mutants remained intact. Together, these data support the notion that the SdeA PDE domain recognises Ub in a manner that is similar to the strategy observed for SdeD, although the interaction is markedly weaker as evidenced by the NMR-titration analysis.

To further address the question of how the ADPR moiety of ADPR-Ub fits in the active-site groove of the PDE domain, we determined the structure of a catalytically inactive SdeD mutant (H67A) in complex with ADPR-Ub. The binding mode of ADPR-Ub is similar to Ub1 with the ADPR moiety nestled in the catalytic groove (Extended Data Fig. 9a–d). ADPR sits atop several invariant residues, including H67A, H189 and E126, and engages in extensive interactions, with a large number of conserved residues within the catalytic groove (Fig. 4a–c, Extended Data Fig. 9e). To test the role of the ADPR-interacting residues within the catalytic groove, we mutated several corresponding residues in SdeA. PDE activity was completely abolished in the H277A, H407A, and E340A mutants, as indicated by the lack of both the Pro-Q staining signal and RAB33B ubiquitination (Fig. 4c, d). The activity of the R413A mutant was substantially impaired, whereas H281A and W394A mutations showed little or no effect on PDE activity.

Based on our results, we propose a two-step reaction mechanism for the transfer of PR-Ub to a substrate (Fig. 4e). In the first step, negatively-charged E340 helps to position R42 of ADPR-Ub and H277. This interaction could enhance the nucleophilicity of H277 through induction. H277 attacks the β-phosphate of ADPR to form a transient phosphoramidate bond with PR-Ub. The presence of this transient intermediate is supported by biochemical evidence reported in an accompanying paper²⁶. The nearby H407 residue functions as a general acid to donate a proton to the α-phosphate of the releasing AMP molecule. The underlying mechanism of this step is similar to that of histidine protein kinases^{27,28}. In the second step, H407 deprotonates the hydroxyl group of a serine residue on the approaching substrate. The activated hydroxyl group then attacks the phosphoryl group to form a stable phosphoserine linkage between the substrate protein and PR-Ub. The protonated E340 then functions as a general acid to protonate H277, thereby regenerating the enzyme to its initial state. Alternatively, if a water molecule serves as the Ub acceptor in the second step, the reaction results in the cleavage of ADPR-Ub to PR-Ub.

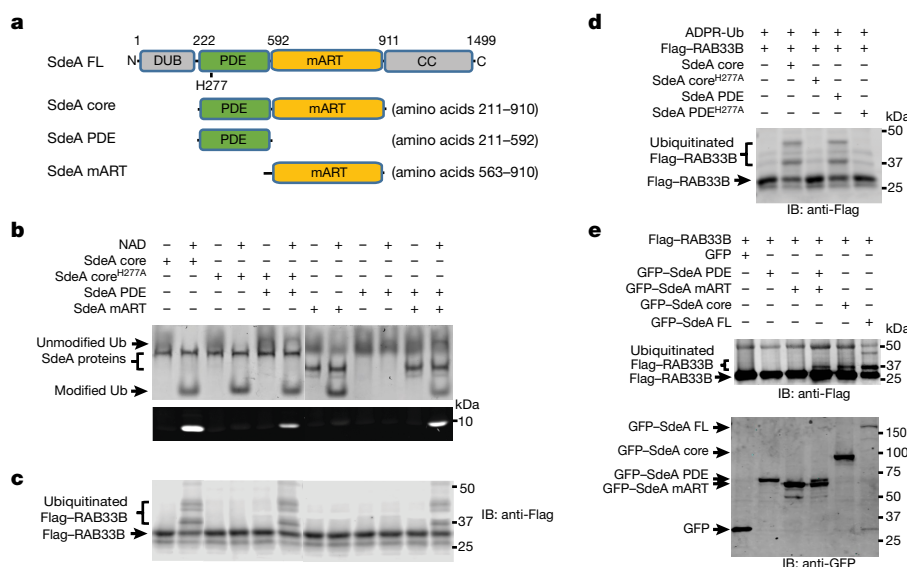


Fig. 2 | ADP-ribosylation of Ub and phosphoribosyl-linked ubiquitination of serine are two independent activities of SdeA.

a, Schematic of SdeA constructs. SdeA has an N-terminal deubiquitinase (DUB) domain, followed by PDE, mART and C-terminal coiled-coil (CC) domains. **b**, In vitro Ub-modification assays. The modification of Ub to ADPR-Ub or PR-Ub was monitored by the band-shift of Ub in native PAGE with Coomassie staining (top). The production of PR-Ub was visualized by SDS-PAGE and phosphoprotein staining with Pro-Q Diamond (bottom). ADPR-Ub and PR-Ub migrate at the same position

on a native gel (labelled as modified Ub), however, only PR-Ub is visible by Pro-Q phosphoprotein stain. **c**, In vitro phosphoribosyl-ubiquitination assay of RAB33B by indicated the SdeA proteins. IB, immunoblot. **d**, In vitro phosphoribosyl-ubiquitination assay of RAB33B in the presence of purified ADPR-Ub. **e**, Intracellular-ubiquitination assays of RAB33B by SdeA. Data shown in **b–d** are representative of four independent experiments. GFP, green fluorescent protein. **e**, Similar results were obtained from three independent experiments. **b–e**, Uncropped gels and blots are shown in Supplementary Fig. 1.

Modification of Ub to yield PR-Ub has not, to our knowledge, been reported in (non-infected) eukaryotes. However, many *Legionella* effector proteins have eukaryotic origins evolutionarily²⁹, raising the

possibility that eukaryotes also harbour an equivalent machinery that may be encoded in multiple polypeptides, as the mART and PDE activities are functionally independent. Future investigation of such

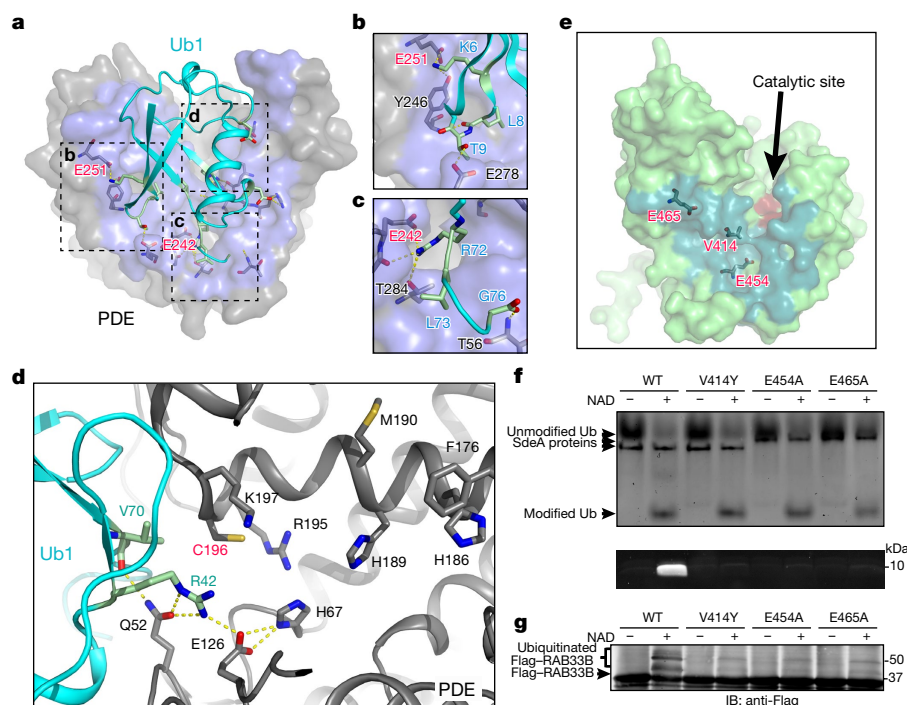


Fig. 3 | The interaction between Ub and the PDE domains of SdeD and SdeA. **a**, Overall view of the binding of Ub (Ub1) with the PDE domain of SdeD. The PDE domain residues within Van der Waals distance of Ub1 are coloured in light blue. Three interacting regions of Ub1 that contact SdeD are marked by dashed outlines. **b–d**, Expanded views of the three Ub1–SdeD interacting regions outlined in **a**. **e**, Surface representation of the PDE domain of SdeA. Ub-binding was modelled the SdeD–Ub1 complex structure and the potential Ub-interacting surface is highlighted

in dark green. Three key residues (E465, E454 and V414) at the potential Ub-interacting interface are shown in stick representation. The PDE active site is shown in red. **f**, **g**, In vitro Ub-modification (**f**) and phosphoribosyl-ubiquitination assays (**g**) of SdeA mutants at the potential Ub interacting interface. The modification of Ub and phosphoribosyl-linked ubiquitination were monitored as described in Fig. 2b, c. Data shown in **f** and **g** are representative of four independent experiments. Uncropped gels and blots are shown in Supplementary Fig. 1. WT, wild type.

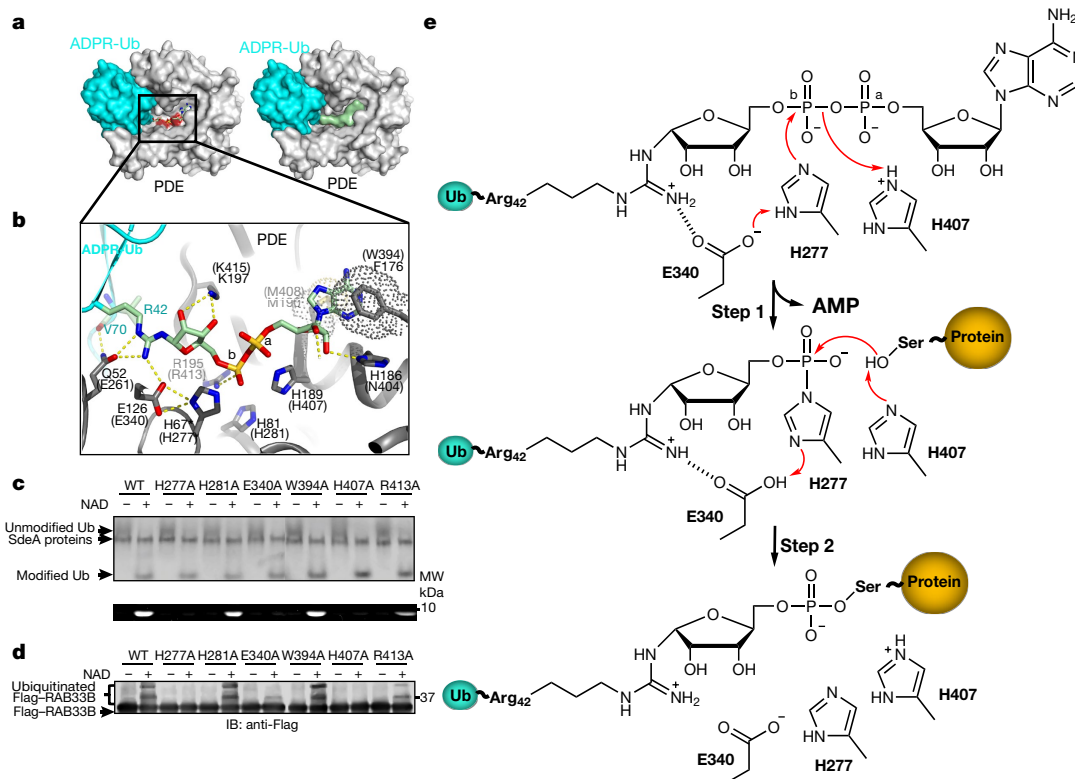


Fig. 4 | Structure of the complex formed by ADPR-Ub and the PDE domain of SdeD. **a**, Surface representation of ADPR-Ub (cyan) in complex with the SdeD PDE domain (grey). The catalytic site is coloured in red. The ADPR moiety is coloured in light green and shown in stick (left) and surface (right) representation. **b**, A detailed interaction of the ADPR moiety with residues of the PDE domain. SdeD residues involved in ADPR-binding are labelled and the corresponding residues in SdeA are labelled in parentheses. In the structure, H67 is substituted with

alanine, but is modelled with histidine and labelled as H67*. **c**, Enzymatic activity analysis of SdeA-core with mutations in conserved residues of the catalytic groove. The modification of Ub was monitored as described in Fig. 2b. **d**, Phosphoribosyl-ubiquitination assay of RAB33B. **e**, A two-step reaction model of phosphoribosyl-linked ubiquitination catalysed by the PDE domain of SdeA. Data shown in **c** and **d** are representative of three independent experiments. Uncropped gels and blots are shown in Supplementary Fig. 1.

a eukaryotic enzyme system will advance our understanding of the versatile Ub code.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0147-6>.

Received: 25 September 2017; Accepted: 18 April 2018;

Published online 23 May 2018.

- Hershko, A., Ciechanover, A. & Varshavsky, A. The ubiquitin system. *Nat. Med.* **6**, 1073–1081 (2000).
- Chen, Z. J. & Sun, L. J. Nonproteolytic functions of ubiquitin in cell signaling. *Mol. Cell* **33**, 275–286 (2009).
- Hurley, J. H. & Stenmark, H. Molecular mechanisms of ubiquitin-dependent membrane traffic. *Annu. Rev. Biophys.* **40**, 119–142 (2011).
- Haglund, K. & Dikic, I. The role of ubiquitination in receptor endocytosis and endosomal sorting. *J. Cell Sci.* **125**, 265–275 (2012).
- Komander, D. & Rape, M. The ubiquitin code. *Annu. Rev. Biochem.* **81**, 203–229 (2012).
- Qiu, J. et al. Ubiquitination independent of E1 and E2 enzymes by bacterial effectors. *Nature* **533**, 120–124 (2016).
- Bhogaraju, S. et al. Phosphoribosylation of ubiquitin promotes serine ubiquitination and impairs conventional ubiquitination. *Cell* **167**, 1636–1649. e1613 (2016).
- Kotewicz, K. M. et al. A single *Legionella* effector catalyzes a multistep ubiquitination pathway to rearrange tubular endoplasmic reticulum for replication. *Cell Host Microbe* **21**, 169–181 (2017).
- Luo, Z. Q. & Isberg, R. R. Multiple substrates of the *Legionella pneumophila* Dot/Icm system identified by interbacterial protein transfer. *Proc. Natl Acad. Sci. USA* **101**, 841–846 (2004).
- Zhou, Y. & Zhu, Y. Diversity of bacterial manipulation of the host ubiquitin pathways. *Cell. Microbiol.* **17**, 26–34 (2015).

- Lin, Y. H. & Machner, M. P. Exploitation of the host cell ubiquitin machinery by microbial effector proteins. *J. Cell Sci.* **130**, 1985–1996 (2017).
- Qiu, J. & Luo, Z. Q. *Legionella* and *Coxiella* effectors: strength in diversity and activity. *Nat. Rev. Microbiol.* **15**, 591–605 (2017).
- Hsu, F. et al. The *Legionella* effector SidC defines a unique family of ubiquitin ligases important for bacterial phagosomal remodeling. *Proc. Natl Acad. Sci. USA* **111**, 10538–10543 (2014).
- Luo, X. et al. Structure of the *Legionella* virulence factor, SidC reveals a unique PI(4)P-specific binding domain essential for its targeting to the bacterial phagosome. *PLoS Pathog.* **11**, e1004965 (2015).
- Price, C. T. et al. Molecular mimicry by an F-box effector of *Legionella pneumophila* hijacks a conserved polyubiquitination machinery within macrophages and protozoa. *PLoS Pathog.* **5**, e1000704 (2009).
- Kubori, T., Hyakutake, A. & Nagai, H. *Legionella* translocates an E3 ubiquitin ligase that has multiple U-boxes with distinct functions. *Mol. Microbiol.* **67**, 1307–1319 (2008).
- Kubori, T., Shinzawa, N., Kanuka, H. & Nagai, H. *Legionella* metaeffector exploits host proteasome to temporally regulate cognate effector. *PLoS Pathog.* **6**, e1001216 (2010).
- Ensminger, A. W. & Isberg, R. R. E3 ubiquitin ligase activity and targeting of BAT3 by multiple *Legionella pneumophila* translocated substrates. *Infect. Immun.* **78**, 3905–3919 (2010).
- Wong, K., Kozlov, G., Zhang, Y. & Gehring, K. Structure of the *Legionella* effector, Ipg1496, suggests a role in nucleotide metabolism. *J. Biol. Chem.* **290**, 24727–24737 (2015).
- Jeong, B. R. et al. Structure function analysis of an ADP-ribosyltransferase type III effector and its RNA-binding target in plant immunity. *J. Biol. Chem.* **286**, 43272–43281 (2011).
- Tsurumura, T. et al. Arginine ADP-ribosylation mechanism based on structural snapshots of iota-toxin and actin complex. *Proc. Natl Acad. Sci. USA* **110**, 4267–4272 (2013).
- Simon, N. C., Aktories, K. & Barbieri, J. T. Novel bacterial ADP-ribosylating toxins: structure and function. *Nat. Rev. Microbiol.* **12**, 599–611 (2014).
- Ashkenazy, H. et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.* **44**, W344–W350 (2016).
- Dong, Y. et al. Structural basis of ubiquitin modification by the *Legionella* effector SdeA. *Nature* <https://doi.org/10.1038/s41586-018-0146-7> (2018).

25. Rinaldo, S. et al. Structural basis of functional diversification of the HD-GYP domain revealed by the *Pseudomonas aeruginosa* PA4781 protein, which displays an unselective bimetallic binding site. *J. Bacteriol.* **197**, 1525–1535 (2015).
26. Kalayil, S. et al. Insights into catalysis and function of phosphoribosyl-linked serine ubiquitination. *Nature* <https://doi.org/10.1038/s41586-018-0145-8> (2018).
27. Klumpp, S. & Krieglstein, J. Phosphorylation and dephosphorylation of histidine residues in proteins. *Eur. J. Biochem.* **269**, 1067–1071 (2002).
28. Stock, A. M., Robinson, V. L. & Goudreau, P. N. Two-component signal transduction. *Annu. Rev. Biochem.* **69**, 183–215 (2000).
29. Burstein, D. et al. Genomic analysis of 38 *Legionella* species identifies large and diverse effector repertoires. *Nat. Genet.* **48**, 167–175 (2016).

Acknowledgements We acknowledge L. Pollack's group for SAXS data collection. This work is supported by National Institute of Health (NIH) grants 5R01GM116964 (Y.M.), R01AI127465 (Z.-Q.L.), R01GM088055 (R.E.K.), 1R01GM098503-05 (P.S.B.), and 1 F32 GM120797 (K.H.R.). The X-ray data were collected at Cornell High Energy Synchrotron Source. CHESS is supported by the NSF and NIH/National Institute of General Medical Sciences (NIGMS) via NSF award DMR-1332208, and the MacCHESS resource is supported by NIH/NIGMS award GM-103485. Some SAXS data were collected at Stanford Synchrotron Radiation Lightsource. Use of the Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory is supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences under Contract No. DE-AC02-76SF00515. The SSRL Structural Molecular Biology

Program is supported by the DOE Office of Biological and Environmental Research, and by the NIH, NIGMS (including P41GM103393).

Reviewer information *Nature* thanks K. Gehring and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions A.A. and D.J.W. performed crystallization, X-ray data collection, structural determination and phosphoribosyl-ubiquitination analysis; X.W. performed protein purification and crystallization; Y.L. and Y.Z. performed the mutagenesis and bacterial infection experiments; Y. Z. and J.Q. performed the ubiquitination-by-co-expression experiments; K.H.R. performed the SAXS experiment; P.S.B. performed the NMR experiment; Z.-Q.L. analysed the data; K.H.R., P.S.B., R.E.K. and Y.M. analysed the data and wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0147-6>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0147-6>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to Y.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Cloning and mutagenesis. DNA fragments encoding the SdeA-core and SdeD($\Delta 1-341$) were amplified from *L. pneumophila* genomic DNA. The PCR products were digested with BamHI and XhoI restriction enzymes and inserted into a pET28a-based vector in-frame with an N-terminal 6 \times His-SUMO tag for protein overexpression in bacteria cells. Amino acid substitutions of SdeA and SdeD were introduced by site-directed mutagenesis using oligonucleotide primer pairs containing the appropriate base changes. The Ub gene was subcloned into a pET21a vector. All constructs were confirmed by DNA sequencing.

Protein expression and purification. Relevant plasmids (containing *Legionella* protein constructs or RAB33B) were transformed into *E. coli* BL21(DE3) cells. Cultures derived from single colonies were grown in Luria-Bertani medium supplemented with 50 $\mu\text{g ml}^{-1}$ kanamycin or 100 $\mu\text{g ml}^{-1}$ ampicillin to mid-log phase. Protein expression was induced with 0.1 mM isopropyl- β -D-thiogalactopyranoside (IPTG) for 12 h at 18 °C. Collected cells were resuspended in a lysis buffer containing 20 mM Tris-HCl (pH 8.0) and 150 mM NaCl and lysed by sonication. Insoluble cellular debris was pelleted by centrifugation at 31,000g for 30 min at 4 °C, and the clarified lysate was incubated with cobalt resin (Gold-Bio) for 1.5 h at 4 °C. Proteins bound to the resin were extensively washed with lysis buffer. The SUMO-specific protease Ulp1 was then added to the resin slurry to release the expressed protein from the His-SUMO tag. Eluted protein samples were further purified by fast protein liquid chromatography (Superdex 16/60, GE Lifesciences) in 150 mM NaCl, 20 mM Tris pH 7.5. Peak fractions were collected, pooled and concentrated. Protocols for Ub expression and purification were adapted from the published literature³⁰. In brief, collected cells were resuspended in 20 mM ammonium acetate, pH 5.1. Cells were lysed by sonication and cell lysate was clarified by centrifugation (31,000g for 30 min). The pH of the clarified lysate was lowered to 4.8 using glacial acetic acid. The decrease in pH caused the lysate to turn milky white (a result of precipitated proteins), and the solution was again centrifuged at 31,000g for 30 min at 4 °C to remove the precipitated protein fraction. The pH of the remaining soluble fraction was adjusted to 5.1 by the addition of NaOH. The soluble fraction was then loaded onto a HiTrap SP cation exchange column (GE Healthcare) in 20 mM ammonium acetate pH 5.1, and eluted in a continuous gradient of 500 mM ammonium acetate pH 5.1. Fractions containing the ubiquitin peak were pooled and further purified using size exclusion chromatography in 150 mM NaCl, 20 mM Tris pH 7.5. Ubiquitin-containing fractions were pooled and concentrated.

To generate ADPR-Ub for both biochemical assays and crystallographic trials, 1 μM SdeA-core^{H277A} (which lacks PDE activity) was incubated with 25 μM Ub and 1 mM NAD⁺ for 1 h at 37 °C. ADPR-Ub was purified by size exclusion chromatography in 150 mM NaCl, 20 mM Tris (pH 7.5).

Protein crystallization. Generally, all protein crystallization screens were performed with a Crystal Phoenix liquid handling robot (Art Robbins Instruments) at room temperature. The crystallization conditions, which yielded the initial crystals from the screen, were further optimized using the hanging-drop vapour diffusion method by mixing 1.5 μl of protein with an equal volume of reservoir solution.

Specifically, for SdeA-core crystallization, SdeA-core protein was concentrated to 12 mg ml⁻¹ and crystallized in 100 mM HEPES pH 7.9, 12% PEG 8000. Thin plate-shaped crystals appeared in about two weeks. For SdeD crystallization, SdeD was concentrated to 14 mg ml⁻¹ and crystallized in 200 mM CaCl₂, 100 mM MES pH 5.5, 18% PEG 6000, and 100 mM DTT. Cube-shaped crystals formed within two to three days. To generate the SdeD-Ub crystals, SdeD($\Delta 1-341$) was mixed with wild-type Ub at a 1:5 molar ratio, with a final SdeD concentration of 8 mg ml⁻¹. Rod-shaped crystals formed in 200 mM NaCl, 100 mM imidazole pH 7.0, and 24% PEG 8000.

We also obtained crystals of a catalytically inactive SdeD^{H67A} mutant with purified ADPR-Ub. However, those crystals diffracted poorly (probably owing to conflicting crystal packing contacts mediated by the ADPR moiety at the Ub2 site). We therefore attempted to crystallize the SdeD PDE domain with a mixture of ADPR-Ub and unmodified Ub in a 1:2:3 molar ratio and a final SdeD concentration of 12 mg ml⁻¹. We expected ADPR-Ub to have a higher affinity for binding at the Ub1 site, allowing unmodified Ub to bind to the Ub2 site to satisfy crystal packing constraints. Rod-shaped crystals appeared in one day in a solution containing 100 mM sodium cacodylate pH 6.7 and 21% PEG 8000. This strategy yielded diffraction quality crystals in which ADPR-Ub is bound at the Ub1 site and unmodified Ub bound at the Ub2 site.

X-ray diffraction data collection and processing. Diffraction datasets for SdeA-core, the SdeD-Ub complex, and the SdeD-Ub-ADPR-Ub complex were collected at Cornell synchrotron light source MacCHESS beamline F1 and datasets for SdeD crystals were collected at the A1 beamline. Before data collection, all crystals were soaked in cryoprotectant solutions containing their respective crystallization con-

dition buffer supplemented with 20% glycerol and flash frozen in a stream of liquid nitrogen. All datasets were indexed, integrated and scaled with HKL-2000³¹.

Structure determination and refinement. The structure of SdeA-core was solved using the single wavelength anomalous dispersion (SAD) method. Before data collection, SdeA-core crystals were soaked in cryoprotectant (0.1 M HEPES pH 7.9, 12% PEG 8000, and 25% (v/v) glycerol) with the addition of 10 mM ethylmercury chloride for 5 min at room temperature. Heavy atom sites were determined and the initial phase was calculated using the program HKL2MAP³². The structure of the PDE domain of SdeD was solved by SAD phasing with selenomethionine-incorporated SdeD crystals. The structures of the SdeD-Ub and SdeD-Ub-ADPR-Ub complexes were solved by molecular replacement with the AMoRe program³³ of the CCP4 suite³⁴, using the apo SdeD structure as the search model. For all datasets, iterative cycles of model building and refinement were carried out with Coot³⁵ and Refmac5³⁶ of the CCP4 suite.

NMR titration analysis. All NMR spectra were collected on a Bruker 500 MHz DMX at 25 °C. Data were processed using NMRPipe³⁷ and analysed using NMRView³⁸. NMR samples were prepared in 25 mM NaPi, 150 mM NaCl buffer at pH 7.0 with 5% (v/v) D₂O. For all NMR experiments, the concentration of ¹⁵N-Ub or ADPR-Ub (in ADPR-Ub only the Ub subunit was isotopically labelled) was maintained at 150 μM . Concentrations of other protein components varied from 35–300 μM . Two independent experiments were collected for the ¹⁵N-Ub + SdeA PDE domain complex. Each experiment used different stocks of Ub and PDE. Four separate samples containing Ub and different concentrations of SdeD were prepared to collect spectra monitoring the interaction between SdeD and Ub (Ub = 150 μM ; SdeD = 37.5, 75, 150 and 300 μM).

SAXS data collection. SAXS experiments were performed on beamline 4-2 at the Stanford Synchrotron Radiation Lightsources (SSRL)³⁹. Concentrated SdeA-core protein samples were buffer exchanged into 20 mM HEPES pH 7.5, 150 mM NaCl, and stored at 4 °C before data collection. Fifty microlitres of SdeA-core (7 mg ml⁻¹) were injected onto a Superdex 200 Increase PC 3.2/30 (GE Healthcare) column in buffer containing 20 mM HEPES pH 7.5, 150 mM NaCl, 5 mM DTT, 0.02% NaN₃, with a flow rate of 0.05 ml min⁻¹ for online SEC-SAXS. Data were collected using a Pilatus3 \times 1 M detector with a 2.5 m sample-to-detector distance and X-ray beam energy of 12.4 keV (wavelength, $\lambda = 1 \text{ \AA}$), with 1-s exposures collected every 5 s. The first 100 images were averaged as buffer scattering data and subtracted from the corresponding protein scattering data. SAXS patterns, the radius of gyration (R_g), the maximal particle dimension (D_{max}), and the pairwise distance distribution histogram ($P(r)$ plot) and Kratky plot were analysed using the ATSAS software suite⁴⁰. The AllosMod-FOXS server was used for the comparison of solution and X-ray structure conformations^{41,42}. The X-ray-determined 'open' structure and modelled 'closed' conformation were used as input structures. AllosMod generated one hundred static structures, using MODELLER⁴³, which were similar to the input X-ray determined (open) or modelled (closed) structures of SdeA-core⁴². Theoretical SAXS profiles were calculated and compared against the raw SAXS data using FOXS rigid-body modelling as previously described⁴¹, with a maximal q value of 0.25. The mean and s.d. in χ^2 amongst the five best-fitting models were examined for fit comparisons.

Computational analysis and graphical presentation of protein sequence and structure. Sequences homologous to SdeA were selected from results generated by the BLAST server (NCBI). Edited sequences were aligned with Clustal Omega⁴⁴ and coloured using the Multiple Align Show online server (<http://www.bioinformatics.org/sms/index.html>). Protein surface conservation was calculated using the online ConSurf server (<http://consurf.tau.ac.il>)²³. All structural figures were generated using PyMOL (The PyMOL Molecular Graphics System, v.1.8, Schrödinger, LLC) except for the difference Fourier electron density map figure (Extended Data Fig. 9e), which was generated in Coot. The electrostatic surface potential is calculated using the APBS program (<http://www.poissonboltzmann.org>). The surface is coloured on the basis of electrostatic potential with positively charged regions in blue (+4 kcal per electron) and negatively charged surfaces in red (−4 kcal per electron).

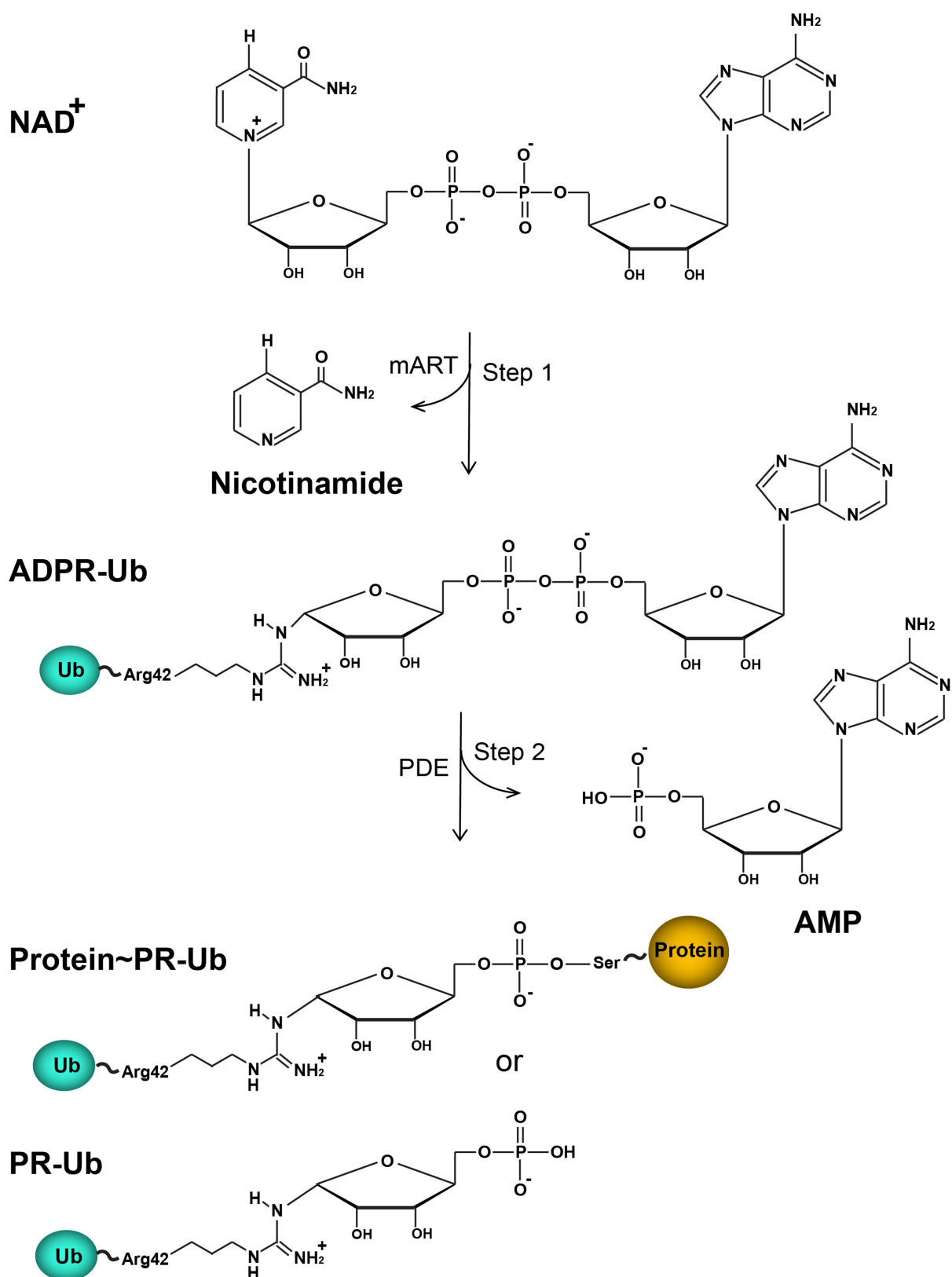
Ubiquitin-modification and RAB33B-ubiquitination assays. Ub-modification reactions were carried out by mixing 1 μM of SdeA-core or SdeA-mART($\Delta 563-910$) with 25 μM ubiquitin in a reaction buffer containing 50 mM NaCl and 50 mM Tris pH 7.5, in the presence or absence of 1 mM NAD⁺. The reactions were incubated for 1 h at 37 °C and reaction products were assessed using both 8% native PAGE and 12% SDS-PAGE. Native gels were stained with Coomassie and SDS-PAGE gels were stained with Pro-Q Diamond phosphoprotein stain (Invitrogen) to assay for PDE activity. ADPR-Ub and PR-Ub migrate to the same position on a native gel (labelled as modified Ub), however, only PR-Ub is visible by Pro-Q phosphoprotein stain owing to its free phosphoryl group⁴⁵. RAB33B ubiquitination reactions were performed with the addition of 4 μM of recombinant Flag-RAB33B to the Ub modification reaction described above. The reaction products were analysed using SDS-PAGE and a western blot with an anti-Flag antibody (Sigma-Aldrich) at a 1:2,500 dilution. To perform the intra-

cellular phosphoribosyl-ubiquitination assay of RAB33B, plasmids expressing Flag–RAB33B, GFP alone or the indicated GFP-tagged SdeA were co-transfected in NIH HEK293T cells. Whole cell lysates were subjected to immunoprecipitation with Flag beads and the products were analysed using anti-Flag western blot. The expression of GFP–SdeA constructs was analysed with an anti-GFP western blot.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

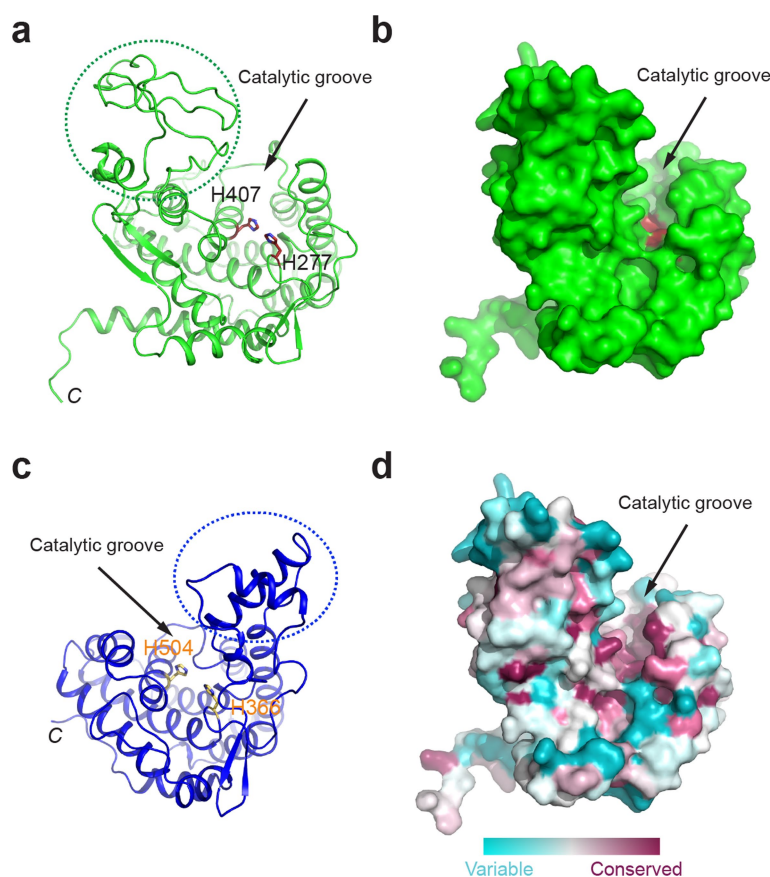
Data availability. Atomic coordinates and structure factors for the reported structures have been deposited into the Protein Data Bank under the accession codes 6B7Q (Hg-bound SdeA), 6B7P (Se-SdeD), 6B7M (SdeD–Ub) and 6B7O (SdeD–Ub–ADPR–Ub). The data supporting the findings of the study are available within the paper and the Extended Data figures and tables. Further data are available from the corresponding author upon reasonable request. The raw images of electrophoreses and western blots can be found in Supplementary Fig. 1.

30. Raasi, S. & Pickart, C. M. Ubiquitin chain synthesis. *Methods Mol. Biol.* **301**, 47–55 (2005).
31. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
32. Pape, T. & Schneider, T. R. HKL2MAP: a graphical user interface for macromolecular phasing with SHELX programs. *J. Appl. Crystallogr.* **37**, 843–844 (2004).
33. Trapani, S. & Navaza, J. AMoRe: classical and modern. *Acta Crystallogr. D* **64**, 11–16 (2008).
34. Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
35. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
36. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).
37. Delaglio, F. et al. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277–293 (1995).
38. Johnson, B. A. & Blevins, R. A. NMR View: A computer program for the visualization and analysis of NMR data. *J. Biomol. NMR* **4**, 603–614 (1994).
39. Smolksy, I. L. et al. Biological small-angle X-ray scattering facility at the Stanford synchrotron radiation laboratory. *J. Appl. Crystallogr.* **40**, s453–s458 (2007).
40. Franke, D. et al. ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J. Appl. Crystallogr.* **50**, 1212–1225 (2017).
41. Schneidman-Duhovny, D., Hammel, M. & Sali, A. FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic Acids Res.* **38**, W540–W544 (2010).
42. Weinkam, P., Pons, J. & Sali, A. Structure-based model of allostery predicts coupling between distant sites. *Proc. Natl Acad. Sci. USA* **109**, 4875–4880 (2012).
43. Martí-Renom, M. A. et al. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325 (2000).
44. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
45. Daniels, C. M., Ong, S. E. & Leung, A. K. Phosphoproteomic approach to characterize protein mono- and poly(ADP-ribosyl)ation sites from cells. *J. Proteome Res.* **13**, 3510–3522 (2014).



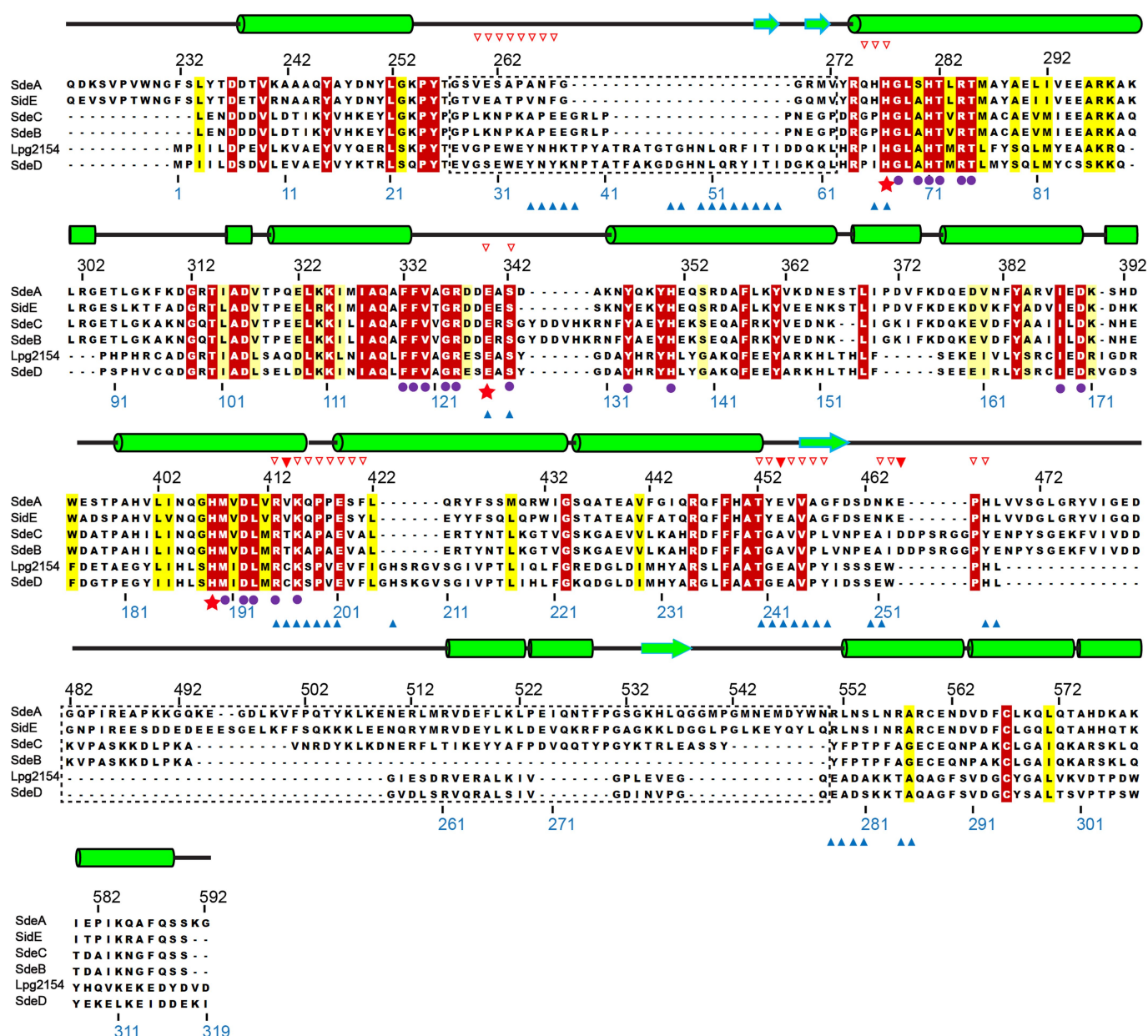
Extended Data Fig. 1 | Chemical structure of phosphoribosyl-linked ubiquitination catalysed by SdeA. Phosphoribosyl-linked ubiquitination catalysed by SdeA involves two enzymatic activities of SdeA. First, using its mART activity, SdeA catalyses the ADP-ribosylation of Ub to generate ADPR-Ub by consuming an NAD⁺ molecule. Second, SdeA catalyses

the conjugation of ADPR-Ub to a serine residue of substrate proteins via its PDE activity to generate protein-PR-Ub and AMP. In the absence of substrate proteins, the PDE domain of SdeA can simply hydrolyse ADPR-Ub to PR-Ub and AMP using a water molecule.



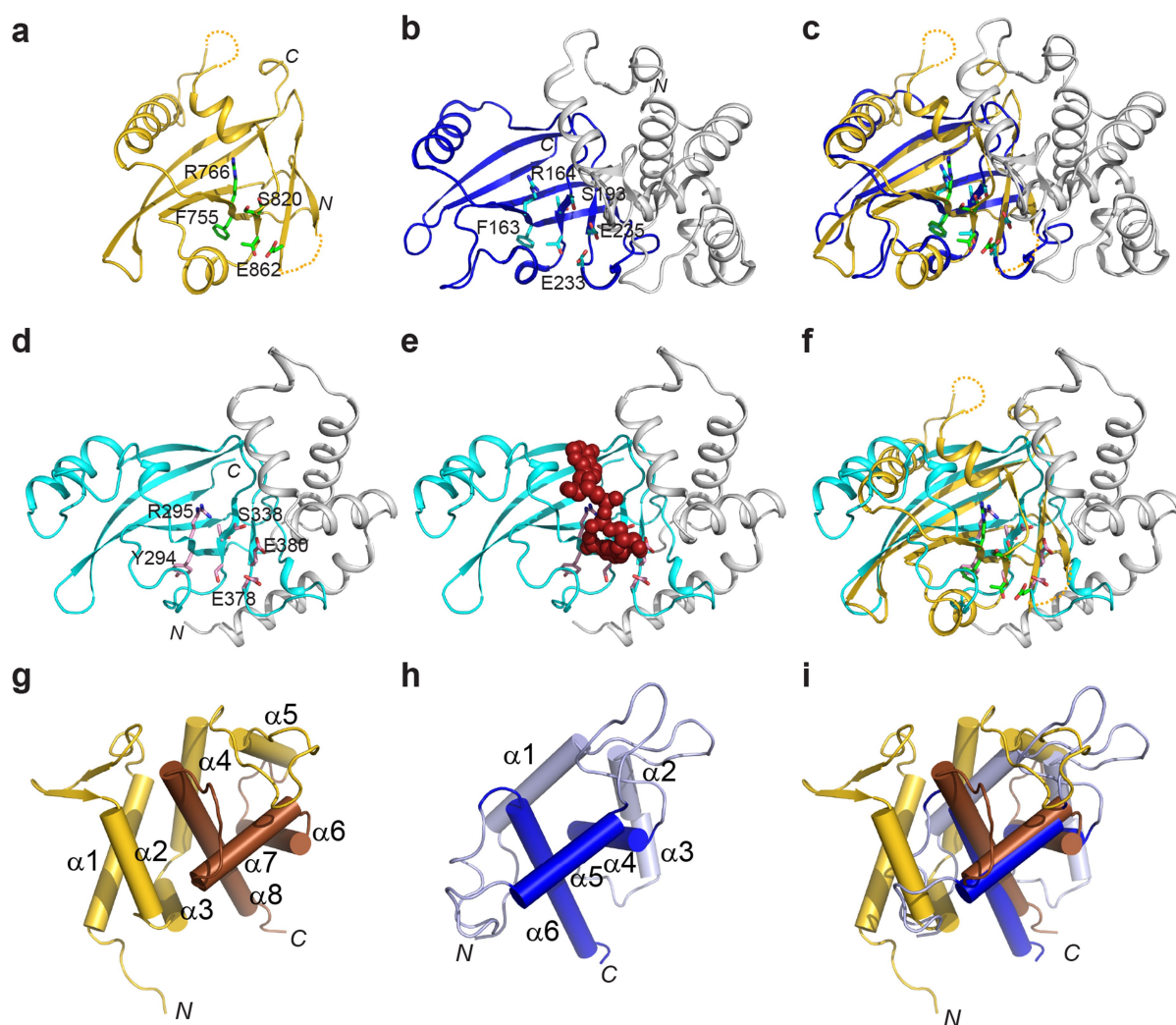
Extended Data Fig. 2 | Structure of the PDE domain of SdeA. **a**, Model of the PDE domain of SdeA in ribbon representation. Two invariant histidine residues (H277 and H407) are shown in stick representation and labelled. **b**, Surface representation of the PDE domain. The two invariant histidine residues (shown in red) are situated at the bottom of a deep groove. **c**, The PDE domain from a *Legionella* effector (lpg1496). Notably the all α -helical structural core of the PDE domains is easy to superimpose onto that of SdeA with a root mean square deviation (r.m.s.d.) of 1.9 Å

over 225 aligned C α atoms. A prominent difference between the two PDE domains is that some loops (indicated by dashed outlines) connecting the α -helices vary both in primary sequence and in length (Extended Data Fig. 3). **d**, Surface residue conservation analysis of the PDE domain. The conservation is calculated using the ConSurf server with the most conserved residues coloured in purple and the least conserved residues in cyan. Note that the catalytic groove is enriched with the most conserved residues.



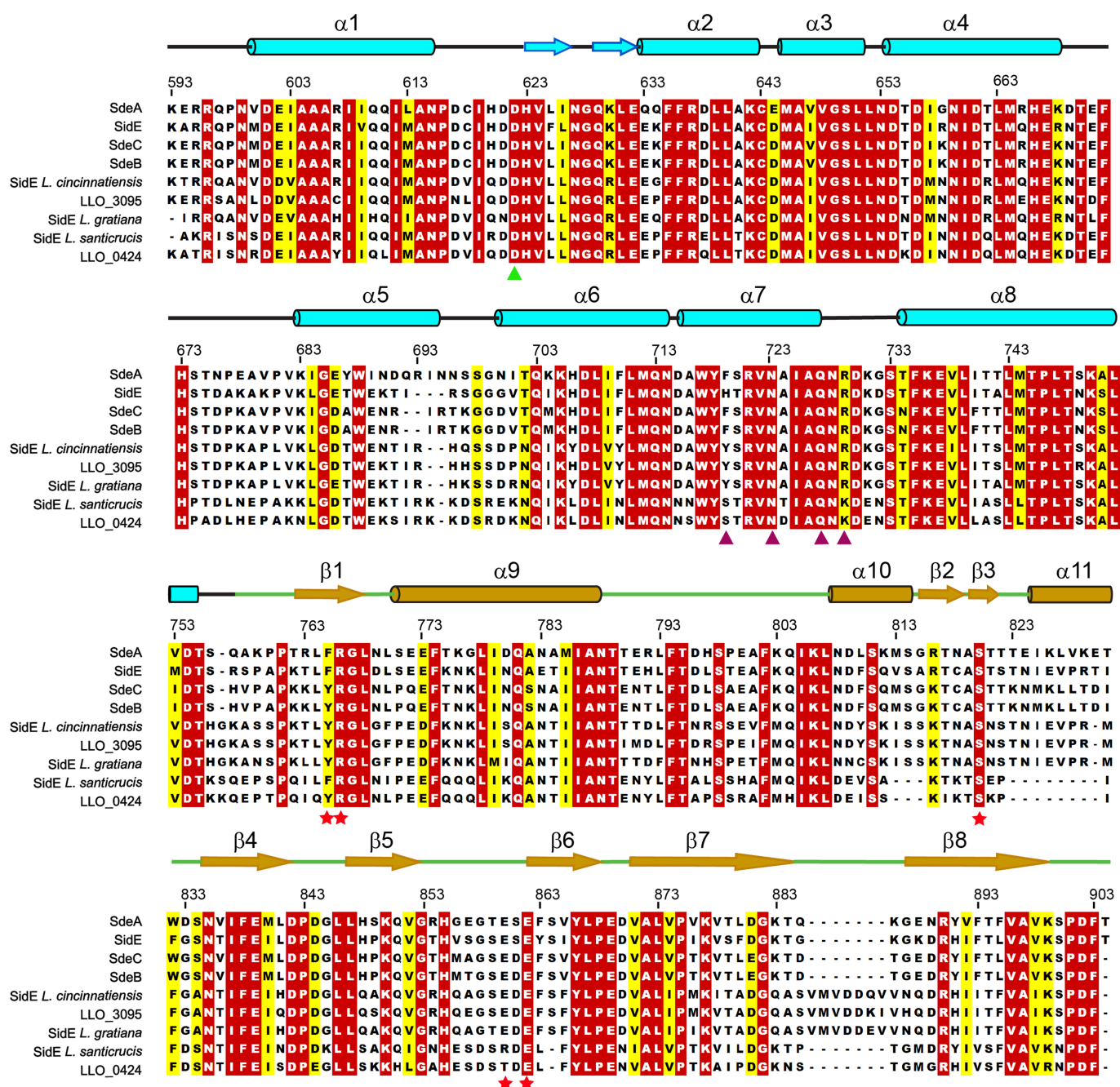
Extended Data Fig. 3 | Multiple sequence alignment of selected PDE domains from the Sde family effectors. Representative sequences corresponding to the PDE domain of SdeA (amino acids 222–502) were aligned using the MultAlin online server (<http://www.bioinformatics.org/sms/index.html>). Secondary structural elements are drawn above the alignment. The numbering for the SdeA sequence is marked on the top of the alignment and the numbering for the SdeD sequence is marked below. Variable loop regions are outlined with dashed squares. Conserved residues located within the catalytic groove are highlighted with purple dots. In particular, three essential catalytic residues (H277, H407 and

E340) are highlighted with red stars below the sequences. SdeD residues that are in close contact with Ub1 (Fig. 3a) are marked by blue triangles at the bottom of the sequences and the predicted Ub1-interacting residues of the PDE domain of SdeA (Fig. 3e) are depicted by red triangles on the top of the sequences. Amongst the potential Ub1-interacting residues, V414, E454 and E465 of SdeA used in mutagenesis studies in Fig. 3f, g are marked with solid red triangles. Entrez database accession numbers are as follows: SdeA, GI: 1064303039; SdeB, GI: 52840489; SdeC, GI: 52842367; SdeD, GI: 52842370; lpg2154, GI: 52842368; and SdeD, GI: 52842717.



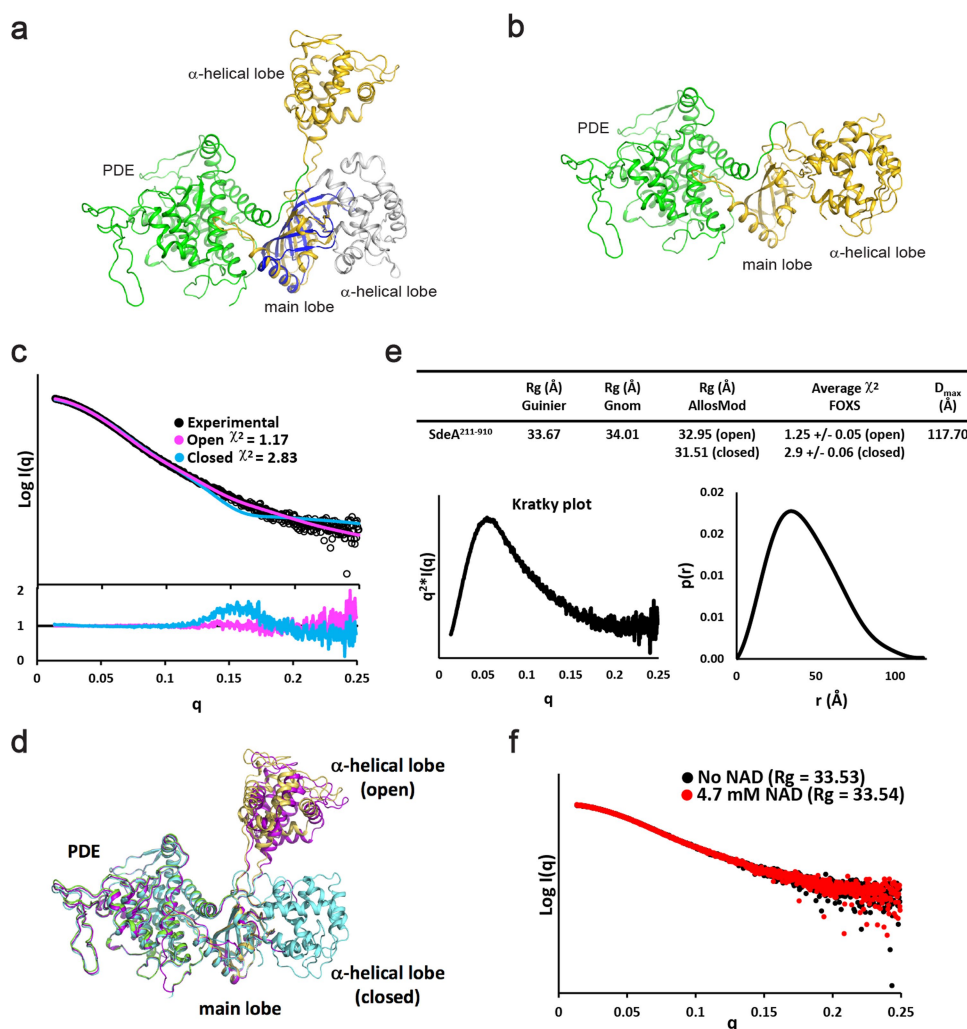
Extended Data Fig. 4 | Structural comparison of the SdeA mART domain with other mART domains from bacterial toxins. **a**, Model of the main lobe of the SdeA mART domain in ribbon representation. The main lobe is composed of two nearly perpendicular β -sheets forming a two-layered β -sandwich core. Residues comprising the three mART catalytic signature motifs: (F/Y)-(R/H), STS and EXE motif are shown in sticks. **b**, HopU1 from *P. syringae* (PDB ID: 3U0J) in ribbon representation. **c**, Structural superimposition of the mART domains from SdeA (gold) and HopU1 (blue). **d**, Iota-toxin from *C. perfringens*

(PDB ID: 4H03). **e**, Iota-toxin in complex with NAD^+ (red spheres). **f**, Structural overlay of the mART domains from SdeA (gold) and Iota-toxin (cyan). **g**, A cartoon diagram of the α -helical lobe of the SdeA mART domain. The α -helical lobe consists of eight α -helices. Three structurally conserved α -helices (α 6–8) are coloured in brown. **h**, A cartoon diagram of the α -helical lobe of HopU1, the three equivalent α -helices (α 4–6) are highlighted in blue. **i**, Structural overlay of the α -helical lobe of SdeA and HopU1.



are close to—the conserved surface patch and are essential for the mART activity (Extended Data Fig. 7), are marked with purple triangles. D622, which is conserved but has no effect on the mART activity is marked with a green triangle. Entrez database accession numbers are as follows: SdeA, GI: 1064303039; SdeE, GI: 52840489; SdeB, GI: 52842367; SdeC, GI: 52842370; SidE *Legionella cincinnatiensis*, GI: 966421657; LLO_3095, GI: 489730495; SidE *Legionella gratiana*, GI: 966468332; SidE *Legionella santacrucis*, GI: 966496250; LLO_0424, GI: 502743808.

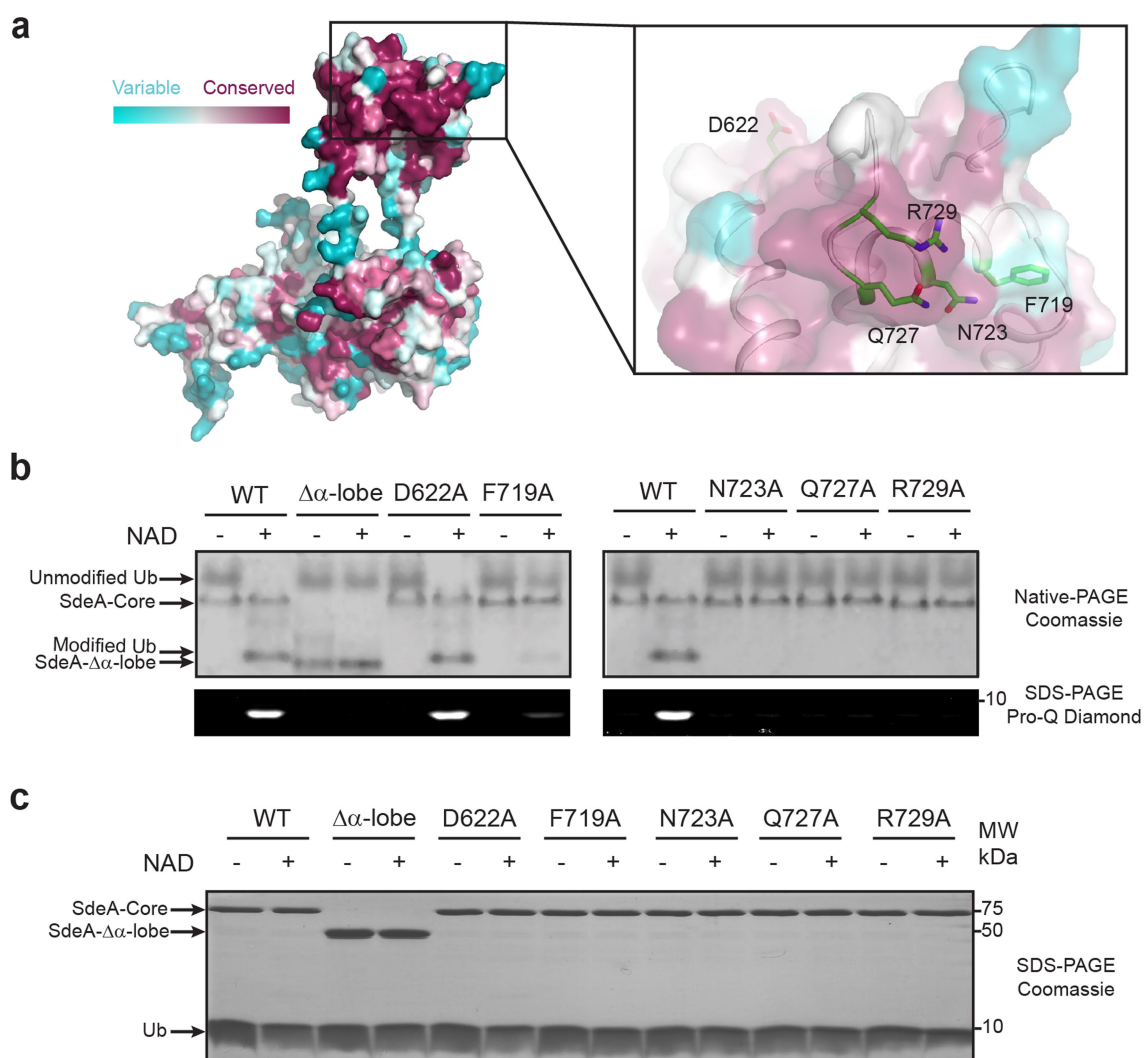
are close to—the conserved surface patch and are essential for the mART activity (Extended Data Fig. 7), are marked with purple triangles. D622, which is conserved but has no effect on the mART activity is marked with a green triangle. Entrez database accession numbers are as follows: SdeA, GI: 1064303039; SdeE, GI: 52840489; SdeB, GI: 52842367; SdeC, GI: 52842370; SidE *Legionella cincinnatiensis*, GI: 966421657; LLO_3095, GI: 489730495; SidE *Legionella gratiana*, GI: 966468332; SidE *Legionella santacrucis*, GI: 966496250; LLO_0424, GI: 502743808.



Extended Data Fig. 6 | The α -helical lobe of SdeA mART domain has an extended conformation compared to other mART proteins.

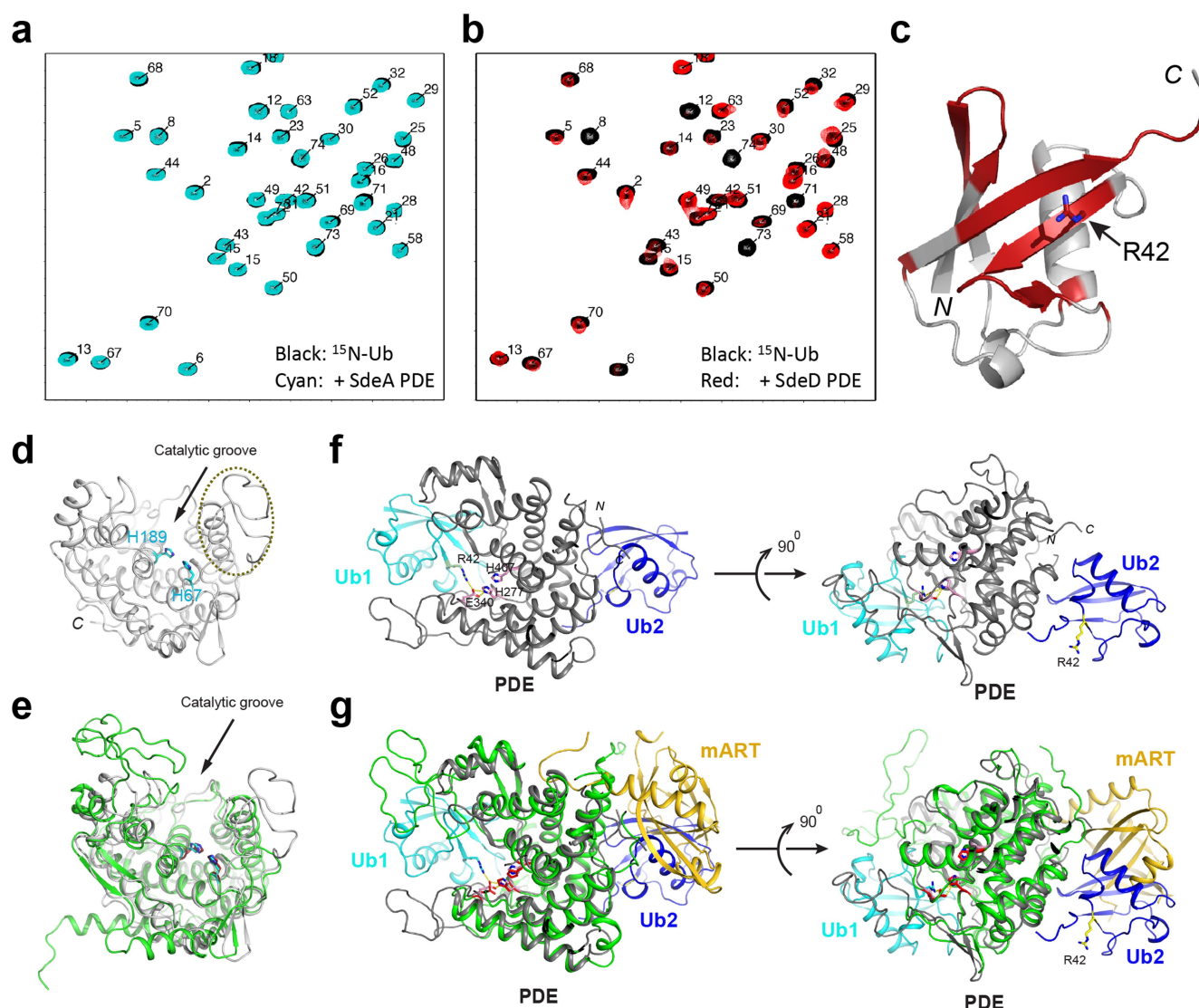
a, Structural superimposition of SdeA onto the HopU1 structure referenced on the main lobe of the mART domain. SdeA is coloured using the same scheme as Fig. 1b. The main lobe of HopU1 is coloured in blue and its α -helical lobe is in grey. The α -helical lobe of the SdeA mART is extended away from the main lobe whereas its counterpart in HopU1 packs in close contact with the main lobe. **b**, Structural model of SdeA with the α -helical lobe in a closed conformation. The positioning of the α -helical lobe was based on a structural overlay of the three structurally conserved α helices identified in all mART domains (Extended Data Fig. 4g–i). **c**, Experimental and theoretical SAXS curves for SdeA-core and the resulting best-fit AllosMod structure for the determined structure (open) and modelled closed conformation, with residual plots shown

below. Best fit χ^2 values are indicated. **d**, Overlay of the determined SdeA-core structure (PDE, green; mART main lobe and α -helical lobe, yellow) and best-fit AllosMod structures for the open (magenta) and closed (cyan) conformations. **e**, Summary of the experimentally derived SAXS parameters for SdeA-core, AllosMod derived best-fit R_g and average FOXS χ^2 for the five best-fitting AllosMod models compared to the experimental SAXS curve. The program Primus was used to calculate the radius of gyration (R_g) and maximum linear dimension (D_{max}). Kratky plot ($I(q)q^2$ versus q), and distance-distribution plot $P(r)$ obtained from GNOM are shown. **f**, Overlay of SdeA-core SAXS curves in the presence of 4.7 mM NAD⁺ ($10\times$ protein concentration), with corresponding Guinier R_g values. Data shown in **c**, **e** and **f** are representative of two biologically independent experiments.



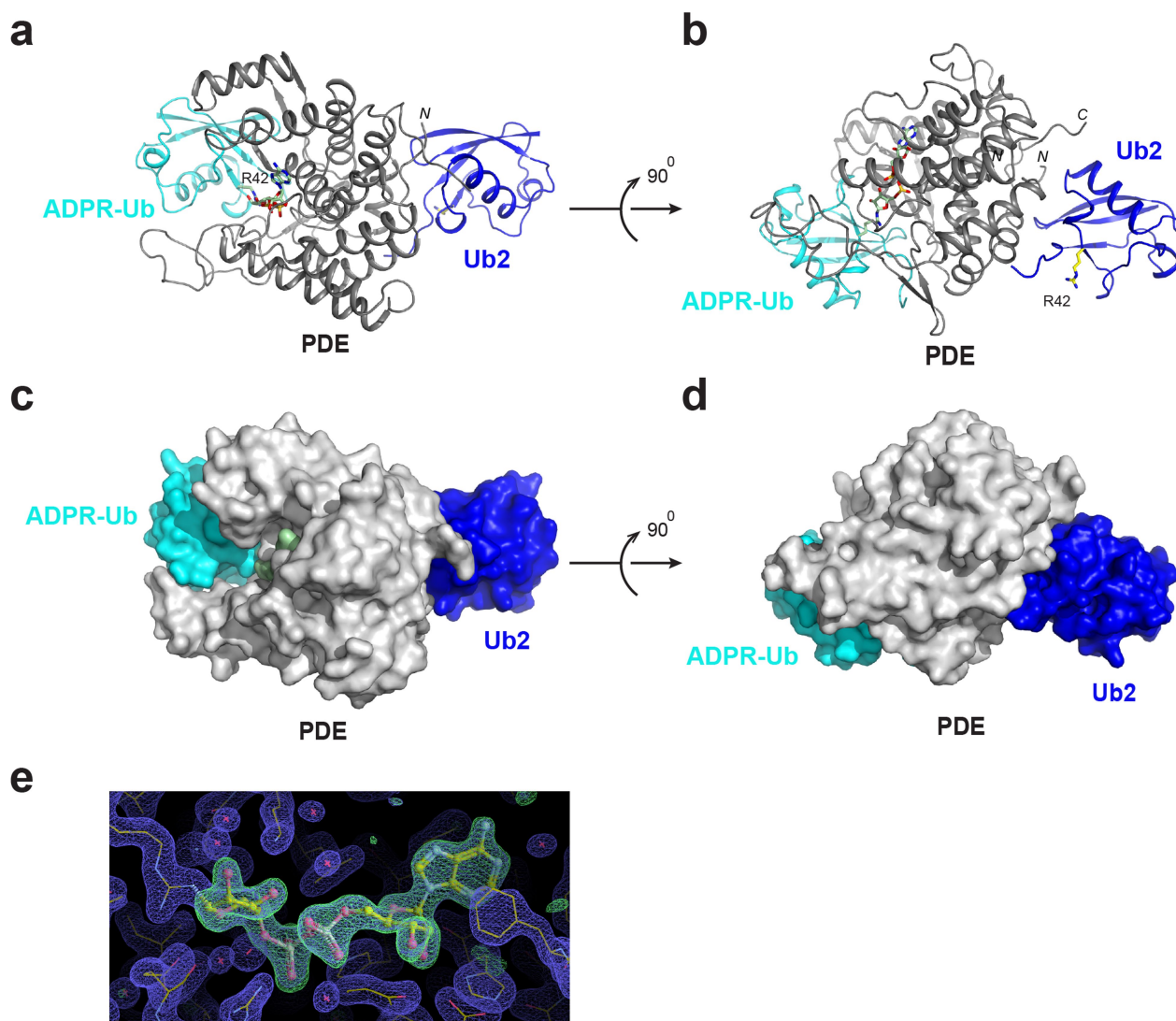
Extended Data Fig. 7 | The α -helical lobe of SdeA mART domain is indispensable for Ub ADP-ribosylation. a, Surface representation of residue conservation of SdeA (the most conserved residues are shown in purple and the least conserved residues in cyan). Surface residue conservation was calculated using the ConSurf server. An expanded view of a surface cluster that consists of the most conserved residues on the α -helical lobe is shown on the right. **b**, Analysis of in vitro ubiquitin-

modification assays by SdeA mutants carrying mutations on the α -helical lobe. The reaction products were analysed using native PAGE with Coomassie blue stain (top) and SDS-PAGE with Pro-Q phosphoprotein stain (bottom). **c**, SDS-PAGE analysis of the proteins in the reaction mixture. Data shown in **b** and **c** are representative of three independent experiments. Uncropped gels are shown in Supplementary Fig. 1.



Extended Data Fig. 8 | The interaction between Ub and the SdeD PDE domain. **a**, NMR ^1H - ^{15}N HSQC TROSY spectral overlay of $150\ \mu\text{M}$ Ub (black) in the presence or absence of $300\ \mu\text{M}$ SdeA PDE domain (cyan). Ub binds very weakly to SdeA as manifested by minimal changes in ^{15}NH peaks of Ub. **b**, Spectral overlay of $150\ \mu\text{M}$ Ub (black) with $75\ \mu\text{M}$ SdeD PDE. Ub binds with higher affinity to SdeD as evidenced by peak broadening and/or disappearance of Ub resonances. **c**, Residues whose resonances are most affected by the presence of SdeD are mapped in red on a cartoon structure of Ub. **d**, PDE domain of SdeD (grey) shown in ribbon representation. Two invariable histidine residues (H67 and H189) are shown in stick representation (cyan). The variable loop unique to SdeD is outlined. **e**, Structural overlay of the PDE domain of SdeD (grey) and the PDE domain of SdeA (green). The overall structures of these two PDE domains are very similar with an r.m.s.d. of $1.73\ \text{\AA}$ over 251 overlaid C α

atoms. **f**, Two orthogonal views of the SdeD PDE domain in complex with two Ub molecules in ribbon representation: Ub1 (cyan) and Ub2 (blue). Ub1 binds at the opening of the PDE catalytic groove with its R42 side chain sticking into the groove. Ub2 binds a region on the opposite side of the catalytic groove. **g**, Structural superimposition of SdeA onto the SdeD PDE–Ub complex referenced on the PDE domain. The PDE domain of SdeA is shown in green and the mART domain is shown in gold. Note that Ub1 shows no conflicting contacts against the superimposed SdeA molecule whereas the Ub2 binding site largely overlaps with the space occupied by the mART domain in SdeA. This analysis suggests that the binding of the PDE domain of SdeD to Ub1 is probably applicable to the PDE domain of SdeA; however, the second Ub-binding site observed in SdeD might not exist in SdeA. Experiments in **a** and **b** were repeated independently two times.



Extended Data Fig. 9 | Crystal structure of the PDE domain of SdeD in complex with ADPR-Ub and Ub. **a**, SdeD PDE domain H67A mutant in complex with both ADPR-Ub and unmodified Ub. The crystal was obtained by mixing the SdeD PDE H67A mutant, ADPR-Ub, and Ub in a 1:2:3 molar ratio (see the ‘Protein crystallization’ section of the Methods for details). The PDE domain is shown in grey, the bound ADPR-Ub is shown in cyan and the unmodified Ub is shown in blue. The unmodified Ub binds a region identical to Ub2 found in the SdeD–Ub complex shown

in Extended Data Fig. 7d. ADPR-Ub binds in a mode that is similar to that of Ub1 in the SdeD–Ub complex with the ADPR moiety fitting into the catalytic groove. **b**, An orthogonal view of **a**. **c**, **d**, Two orthogonal views of the complex shown in **a** in surface representation. Note that the ADPR-moiety shown in light green fits deeply into the catalytic groove. **e**, The density was generated by refinement against the structural model without the ADPR portion. The $F_o - F_c$ difference map is shown in green and contoured at 1σ .

Extended Data Table 1 | X-ray data collection and structural refinement statistics.

	SdeA	SdeD	SdeD-Ub	SdeD-ADPRUB-Ub
Data collection				
Synchrotron beam lines	MCCHES F1	MCCHES A1	MCCHES F1	MCCHES F1
Wavelength (Å)	0.9789	0.68	0.9789	0.9789
Space group	P2 ₁	R3	P2 ₁	P2 ₁
Cell dimensions				
<i>a</i> , <i>b</i> , <i>c</i> (Å)	69.8, 80.6, 85.6	154.4, 154.4, 89.6	64.8, 58.6, 74.1	64.7, 58.8, 75.1
α , β , γ (°)	90, 109.8, 90	90, 120, 90	90, 114.6, 90	90, 114.2, 90
Maximum resolution (Å)	2.2	1.51	1.73	1.88
Observed reflections	61,395	634,900	363,307	281,813
Unique reflections	18,728	124,885	108,100	43,941
Completeness (%) [*]	99.3	99.5	99.4	100
Redundancy [*]	3.4(3.3)	5.1(2.9)	3.4(2.2)	6.4(5.9)
$\langle I \rangle / \langle \sigma I \rangle$ [*]	7.98 (0.87)	29.2 (1.52)	25.4 (1.54)	19.28 (1.18)
R _{sym} (%) [*]	0.122(0.759)	0.07(0.622)	0.078(0.798)	0.093(1.105)
Refinement				
Resolution (Å) [*]	80.51(2.20)	77.174(1.51)	67.36(1.70)	68.93(1.85)
R _{crys} / R _{free} (%) [*]	0.192/0.241	0.167/0.195	0.172/0.28	0.210/0.249
No. atoms				
Protein	5338	4932	3721	3765
Ligand/ion	10	--	--	--
Water	170	448	228	211
<i>B</i> -factors				
Protein	49.728	24.998	28.036	33.327
Ligand/ion	61.563	--	--	--
Water	46.001	31.136	30.525	34.580
R.m.s deviations				
Bond length (Å)	0.023	0.027	0.03	0.028
Bond angles (°)	2.29	2.47	2.47	2.43

^{*}Values in parentheses are for the highest-resolution shell.

Insights into catalysis and function of phosphoribosyl-linked serine ubiquitination

Sissy Kalayil^{1,2,5}, Sagar Bhogaraju^{1,2,5}, Florian Bonn¹, Donghyuk Shin^{1,2}, Yaobin Liu^{1,2}, Ninghai Gan³, Jérôme Basquin⁴, Paolo Grumati¹, Zhao-Qing Luo³ & Ivan Dikic^{1,2*}

Conventional ubiquitination regulates key cellular processes by catalysing the ATP-dependent formation of an isopeptide bond between ubiquitin (Ub) and primary amines in substrate proteins¹. Recently, the SdeA family of bacterial effector proteins (SdeA, SdeB, SdeC and SdeE) from pathogenic *Legionella pneumophila* were shown to use NAD⁺ to mediate phosphoribosyl-linked ubiquitination of serine residues in host proteins^{2,3}. However, the molecular architecture of the catalytic platform that enables this complex multistep process remains unknown. Here we describe the structure of the catalytic core of SdeA, comprising mono-ADP-ribosyltransferase (mART) and phosphodiesterase (PDE) domains, and shed light on the activity of two distinct catalytic sites for serine ubiquitination. The mART catalytic site is composed of an α -helical lobe (AHL) that, together with the mART core, creates a chamber for NAD⁺ binding and ADP-ribosylation of ubiquitin. The catalytic site in the PDE domain cleaves ADP-ribosylated ubiquitin to phosphoribosyl ubiquitin (PR-Ub) and mediates a two-step PR-Ub transfer reaction: first to a catalytic histidine 277 (forming a transient SdeA H277-PR-Ub intermediate) and subsequently to a serine residue in host proteins. Structural analysis revealed a substrate binding cleft in the PDE domain, juxtaposed with the catalytic site, that is essential for positioning serines for

ubiquitination. Using degenerate substrate peptides and newly identified ubiquitination sites in RTN4B, we show that disordered polypeptides with hydrophobic residues surrounding the target serine residues are preferred substrates for SdeA ubiquitination. Infection studies with *L. pneumophila* expressing substrate-binding mutants of SdeA revealed that substrate ubiquitination, rather than modification of the cellular ubiquitin pool, determines the pathophysiological effect of SdeA during acute bacterial infection.

To understand the mode of ubiquitination by SdeA, we sought structural insights into the function of this enzyme. First, we identified SdeA residues 213 to 907 (SdeA₂₁₃₋₉₀₇), comprising both PDE and mART domains, as the minimal stable fragment that can ubiquitinate the known SdeA substrate Rab33b^{2,3}, albeit less efficiently than full-length SdeA (SdeA_{FL}) (Extended Data Fig. 1a, b). We crystallized SdeA₂₁₃₋₉₀₇ and determined its structure at 2.8 Å (Supplementary Table 1, Supplementary Information). In the structure, each asymmetric unit contained one molecule of SdeA₂₁₃₋₉₀₇ comprising three distinct domains (Fig. 1a). The PDE domain spans residues 222–593 and is α -helical. Structure comparison analysis revealed that the PDE domain of SdeA is most similar to that of the *Legionella* effector protein lpg1496 (PDB: 5BU2) (root mean squared deviation (r.m.s.d.) of 2.3 Å over 239 C α atoms)⁴. The closest structural mammalian homologue

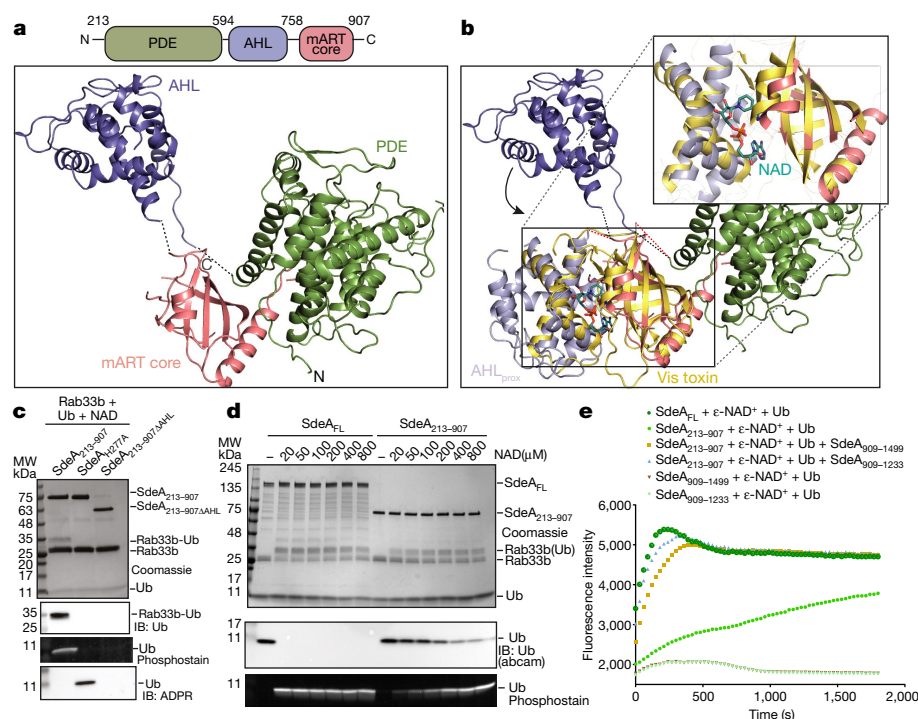


Fig. 1 | Crystal structure of the SdeA catalytic core. **a**, Crystal structure and domain organization of SdeA₂₁₃₋₉₀₇ stable fragment identified by limited proteolysis. Loops connecting the AHL to PDE and mART domains are partially disordered in the crystal structure and are depicted as dotted lines for clarity. **b**, Superimposition of mART core and SdeA AHL in proximal orientation with the mART core of Vis toxin from *V. splendidus* (PDB: 4Y1W). Inset shows the NAD⁺ binding pocket between the mART core and AHL in proximal orientation. **c**, In vitro Rab33b ubiquitination assays comparing the activity of SdeA₂₁₃₋₉₀₇ and SdeA₂₁₃₋₉₀₇ΔAHL. ProQ-Diamond phosphatase was used to monitor PR-Ub. **d**, NAD⁺ sensitivity of SdeA_{FL} and SdeA₂₁₃₋₉₀₇. Abcam ubiquitin (Ub) antibody was used to monitor the levels of unmodified ubiquitin. **e**, ε-NAD⁺ hydrolysis assay with various constructs of SdeA. Experiments were repeated independently three times with similar results (c–e). For gel source data, see Supplementary Fig. 1.

¹Institute of Biochemistry II, Goethe University Frankfurt - Medical Faculty, University Hospital, Frankfurt am Main, Germany. ²Buchmann Institute for Molecular Life Sciences, Goethe University Frankfurt, Frankfurt am Main, Germany. ³Purdue Institute of Immunology, Inflammation and Infectious Diseases and Department of Biological Sciences, Purdue University, West Lafayette, IN, USA. ⁴Max Planck Institute of Biochemistry, Department of Structural Cell Biology, Martinsried, Germany. ⁵These authors contributed equally: Sissy Kalayil, Sagar Bhogaraju. *e-mail: ivan.dikic@biochem2.de

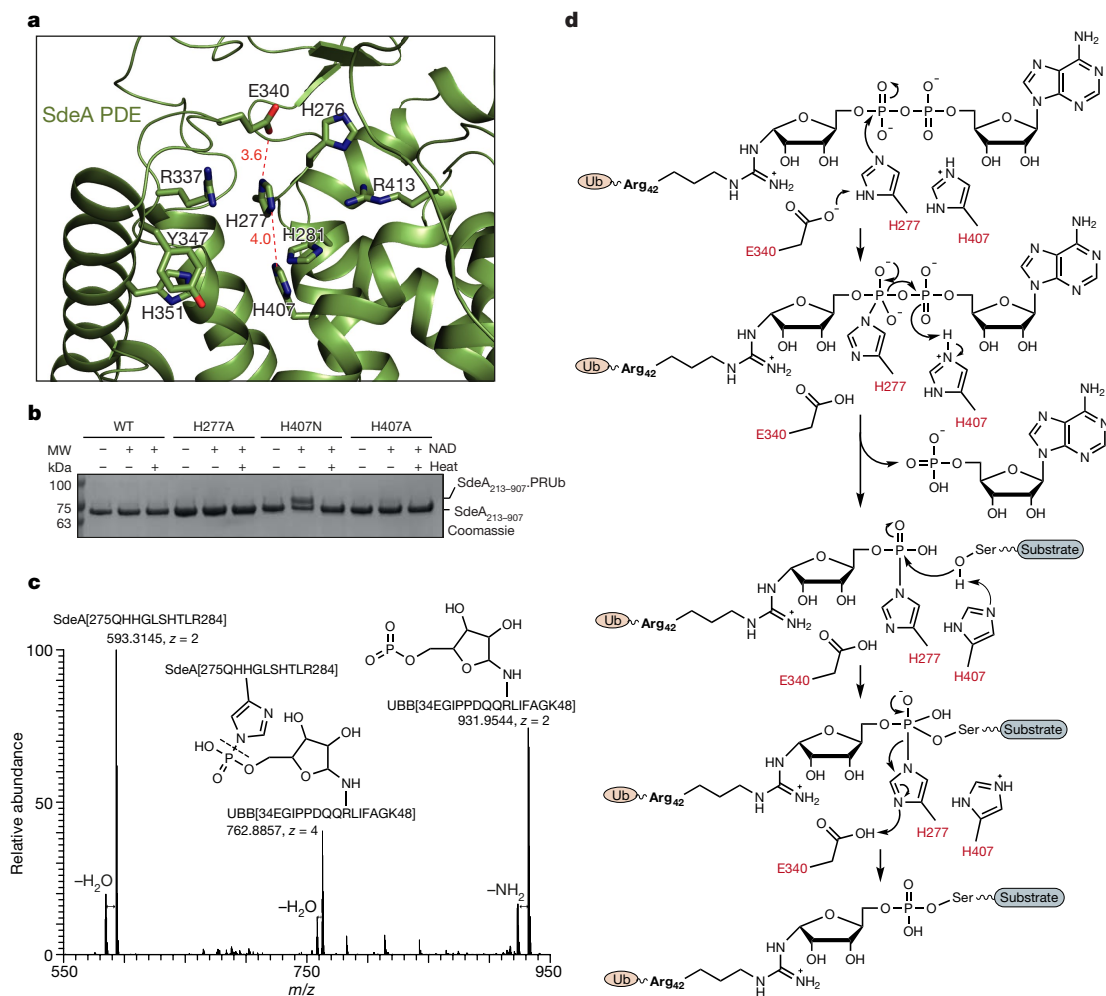


Fig. 2 | Catalytic mechanism of SdeA PDE domain. a, Active site of SdeA PDE domain depicting residues that are important for catalysis. Distances between important catalytic residues are indicated in Ångströms. **b**, In vitro ubiquitination assays with PDE histidine mutations. This experiment was repeated independently three times with similar results. **c**, Stabilized intermediate was analysed by targeted liquid chromatography with tandem mass spectrometry (LC-MS/MS) after tryptic digestion. We used low-energy HCD to target the phosphoramidate bond specifically for partial

fragmentation, creating—besides the intact precursor—marker ions for the tryptic peptides of the active site of the SdeA PDE domain and PR-Ub. This experiment was repeated independently twice with similar results. **d**, Proposed catalytic cycle in SdeA PDE domain mediated by E340, H277 and H407. Electron transfer is indicated by curved arrows. A detailed description of the mechanism can be found in the Supplementary Information. For gel source data, see Supplementary Fig. 1.

of the SdeA PDE domain is human SAMHD1, a dNTP hydrolase with functions in the innate immune response (r.m.s.d. of 4.1 Å over 165 C α atoms)⁵. The mART domain is situated at the C terminus (residues 594–907) and comprises two distinct and spatially separated lobes, the α -helical lobe (AHL, residues 594–758) and the mART core (residues 759–907). The mART core interacts strongly with the PDE domain and is composed mostly of β -strands with a couple of α -helices. Unexpectedly, in our crystal structure, the AHL has no physical proximity to the mART core; this contrasts with the structures of other bacterial ADP-ribosylating enzymes, in which the AHL is an integral part of the mART domain and contributes to NAD⁺ binding and ADP-ribosylation of the substrate⁶. The solution structure of SdeA₂₁₃₋₉₀₇ that was determined using small-angle X-ray scattering (SAXS) revealed a similar orientation of AHL in solution as seen in the crystal (Extended Data Fig. 1c, Supplementary Table 2, Supplementary Information). Superimposition of the AHLs of SdeA and Vis toxin, a bacterial ADP-ribosyl transferase from *Vibrio splendidus* (PDB: 4Y1W), revealed a proximal conformation of the AHL that differed substantially from that seen in the crystal structure (Fig. 1b). We hypothesize that the AHL of SdeA₂₁₃₋₉₀₇ could transiently adopt a conformation proximal to the mART core for NAD⁺ binding and processing (Fig. 1b). Consistent with this hypothesis, deletion of the AHL (residues 599–758) led to a

complete loss of ADP-ribosylation of ubiquitin and ϵ -NAD⁺ hydrolysis⁷ (Fig. 1c, Extended Data Fig. 2a). Mutation of residues in the two flexible loops flanking the AHL affected substrate ubiquitination in SdeA₂₁₃₋₉₀₇ but not in SdeA_{FL}, suggesting that the dynamic conformational shift of AHL occurs only in the context of SdeA₂₁₃₋₉₀₇, while the position of AHL in SdeA_{FL} is fixed in the proximal, active form by the C-terminal region (CTR, residues 909–1499) (Extended Data Fig. 2b, c). Accordingly, SdeA_{FL} exhibited much greater NAD⁺ sensitivity in our in vitro ubiquitination experiments, resulting in complete modification of 10 μ M ubiquitin with 20 μ M NAD⁺, whereas the activity of SdeA₂₁₃₋₉₀₇ gradually increased with increasing NAD⁺ concentration (Fig. 1d). Similarly, SdeA_{FL} exhibited a marked increase in activity compared to SdeA₂₁₃₋₉₀₇ with respect to the ϵ -NAD⁺ hydrolysis kinetics measured in vitro (Fig. 1e). SdeA₂₁₃₋₉₀₇ did not detectably ubiquitinate Rab33b in HEK293T cells, perhaps owing to insufficient cellular NAD⁺ concentration (Extended Data Fig. 2d). Moreover, limited proteolysis experiments with SdeA constructs containing different C-terminal extensions revealed that the construct ending at residue 1233 is indigestible, whereas shorter constructs collapse to SdeA₂₁₃₋₉₀₇, indicating that the CTR induces a compact or closed state of the SdeA structure (Extended Data Fig. 2e). Mixing purified CTR (residues 909–1499) or shorter CTR (residues 909–1233) with SdeA₂₁₃₋₉₀₇ increased the

catalytic activity of SdeA_{213–907} to the same level as SdeA_{FL} (Fig. 1e). These results are consistent with the crystal structure of a longer construct of SdeA in which CTR stabilizes the proximal orientation of AHL⁸. Detailed analysis of the structure of the mART domain and its interaction with the PDE domain is presented in the Supplementary Information (Extended Data Figs. 3, 4).

The PDE domain of SdeA hydrolyses ADP-ribosylated ubiquitin (ADPR-Ub) and catalyses the transfer of ubiquitin to serine residues of the substrate protein via a phosphoribose linker³. The catalytic pocket of PDE in SdeA is lined by several conserved histidines (Fig. 2a) and mutation of H277 or H407 has been shown to abolish the activity of the PDE domain⁹. We hypothesized that the PR-ubiquitination reaction by SdeA might take place via a transient intermediate involving covalent attachment of phosphate to a catalytic histidine residue¹⁰. Phosphohistidine intermediates are difficult to observe owing to their extreme lability^{11, 12}. Therefore, we introduced a few key mutations into the catalytic pocket of the PDE domain to stabilize a potential intermediate. Unexpectedly, we observed a SdeA-Ub intermediate that was sensitive to heat treatment only in the H407N mutant (Fig. 2b, Extended Data Fig. 5a). Using mass-spectrometry analysis of the tryptic digest of the intermediate reaction, we could identify an ion with the exact mass of ubiquitin (Ub) 34–48 bridged to SdeA 275–284 by phosphoribose. We applied low-energy HCD (high-energy collisional dissociation) fragmentation to specifically cleave the phosphoramidate bond, and confirmed that the generated fragment ion corresponded to SdeA 275–284 and phosphoribosylated Ub 34–48 (Fig. 2c). This analysis revealed that H277 of SdeA is linked by phosphoribose to Ub through a phosphoramidate bond. We further validated the identity of the histidine-bridged intermediate by high-energy fragmentation generated ion series of the peptide backbones (Extended Data Fig. 5b). Using rhodamine-labelled and haemagglutinin-labelled ubiquitin, we observed that both ubiquitin variants attached to SdeA_{213–907}(H407N) in a heat-dependent manner (Extended Data Fig. 5c, d). Upon heating, the levels of intermediate decreased with a concomitant increase of PR-Ub, indicating that PR-Ub is attached to a catalytic histidine of SdeA (Extended Data Fig. 5e). Notably, residue E340 of SdeA forms a hydrogen bond with H277, potentially activating this histidine to be a strong nucleophile¹³ (Fig. 2a). Both double mutants (H277A/H407N and E340A/H407N) failed to form the intermediate, supporting our notion that H277 is the intermediate-forming residue and that E340 has a critical role in the activation of H277 (Extended Data Fig. 5f). Based on the observed stabilization of the intermediate in H407 mutants, we propose a general role for H407 in orienting and activating a water molecule or the substrate serine for a nucleophilic attack on H277-PR-Ub (Fig. 2d). Accordingly, SdeA_{213–907}(H407N) is defective in substrate ubiquitination and only partially active in producing PR-Ub (Extended Data Fig. 6a). We propose a two-step phosphoryl transfer reaction scheme for the PDE-mediated PR-ubiquitination of substrates (Fig. 2d). This model is also supported by the crystal structure of ADPR-Ub in complex with the PDE domain of SdeD¹⁴. In addition to revealing critical roles for the two histidine residues, the structure of PDE also showed that side chains of Y347 and R413 are inserted into the catalytic centre and could be potentially involved in either the phosphoryl transfer activity or the binding of ADPR-Ub to SdeA (Fig. 2a). Mutating these residues into alanine inhibited PR-ubiquitination by SdeA (Extended Data Fig. 6b, c).

To gain insights into substrate recognition by SdeA, we set out to identify ubiquitination sites within the recently described SdeA substrate RTN4B using a mass-spectrometry-based approach³. We found two ubiquitination sites in the cytoplasmic part of RTN4B, where each site contained two serine residues that are targeted by SdeA for ubiquitination (Extended Data Fig. 7a, b). RTN4B peptides of about 13 residues containing the target serine residues served as substrates for SdeA (Fig. 3a). Alignment of all peptide sequences containing known SdeA target serines using Seq2logo server¹⁵ generated a sequence motif (Fig. 3b, Extended Data Fig. 7c), in which the target serine is in the vicinity of hydrophobic residues and is flanked by proline residues.

Furthermore, we produced 47 degenerate peptides based on one of the RTN4B ubiquitination sites and performed ubiquitination assays with all of them (Supplementary Table 3, Supplementary Information, Fig. 3c). Individual peptide sequences and their fractional activity were given as inputs for NNalign server to generate a sequence motif¹⁶. The resulting sequence motif confirms the importance of hydrophobic residues surrounding the target serine sites for SdeA ubiquitination (Fig. 3d). Notably, we have identified an AMP analogue, adenosine 5'-O-thiomonophosphate (AMPS), that acts as a low-affinity inhibitor of substrate ubiquitination by the SdeA PDE domain by affecting substrate positioning (Supplementary Information, Extended Data Fig. 8). AMPS also inhibited the ubiquitination activity of SdeC, a paralogue of SdeA (Extended Data Fig. 8g), indicating the potential implications of this lead compound in developing novel inhibitors against this class of *Legionella* toxins.

Further inspection of the SdeA PDE domain structure revealed the existence of a cleft on the surface that leads to the active site of the PDE containing the catalytic residue H277 (Fig. 4a). We hypothesized that this cleft could bind and position polypeptides containing substrate serines. We identified several SdeA cleft mutants (Fig. 4b) that are associated with a substantial decrease in substrate ubiquitination, while showing negligible effect on phosphoribosylation of ubiquitin (Fig. 4c, Extended Data Fig. 9a). Among the mutants, M408A, L411A and M408A/L411A had the biggest effect on Rab33b ubiquitination in both SdeA_{213–907} and SdeA_{FL}. These mutations also affected ubiquitination of RTN4B, indicating that SdeA recognizes multiple structurally different substrates via this region (Fig. 4d).

The identification of substrate-binding mutants of SdeA enabled us to investigate which of its two functions (phosphoribosylation of ubiquitin or substrate ubiquitination³) is physiologically relevant. The *L. pneumophila* Δ sidEs mutant² was complemented with either wild-type SdeA_{FL} or various substrate-binding mutants (Extended Data Fig. 9b). Both the M408A and M408A/L411A mutants lacked the ability to complement the growth of the bacterial strain lacking the *sidE* effector family in the amoeba host *Dictyostelium discoideum* (Fig. 4e, Extended Data Fig. 9c) and also failed to restore the *sidEs* mutant in the recruitment of RTN4 to the *Legionella*-containing vacuole during infection in primary mouse macrophages (Fig. 4f, Extended Data Fig. 9d). Together, our results indicate that targeting of specific

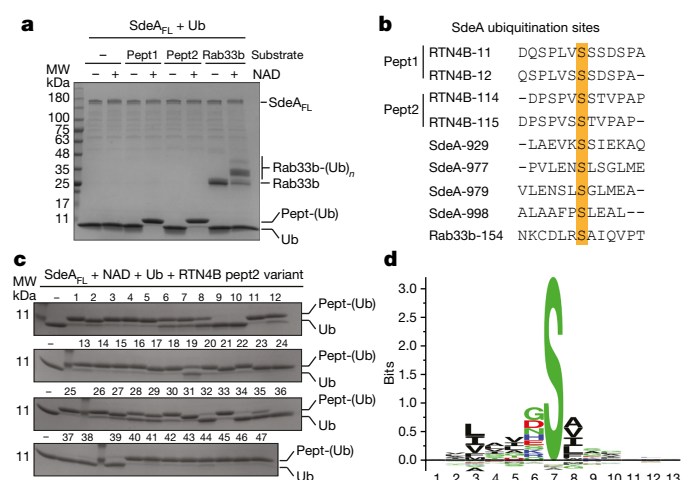


Fig. 3 | Substrate recognition by SdeA. **a**, In vitro ubiquitination of RTN4B peptides (Pept1 and Pept2) and Rab33b by SdeA_{FL}. **b**, Alignment of target serine sequences of SdeA identified so far. **c**, In vitro ubiquitination of 47 degenerate peptides designed with RTN4B substrate peptide as template. **d**, Sequence motif generated by NNalign software, resulting from analysis of in vitro ubiquitination data of the peptides. Experiments were repeated independently twice with similar results (**a**, **c**). For gel source data, see Supplementary Fig. 1.

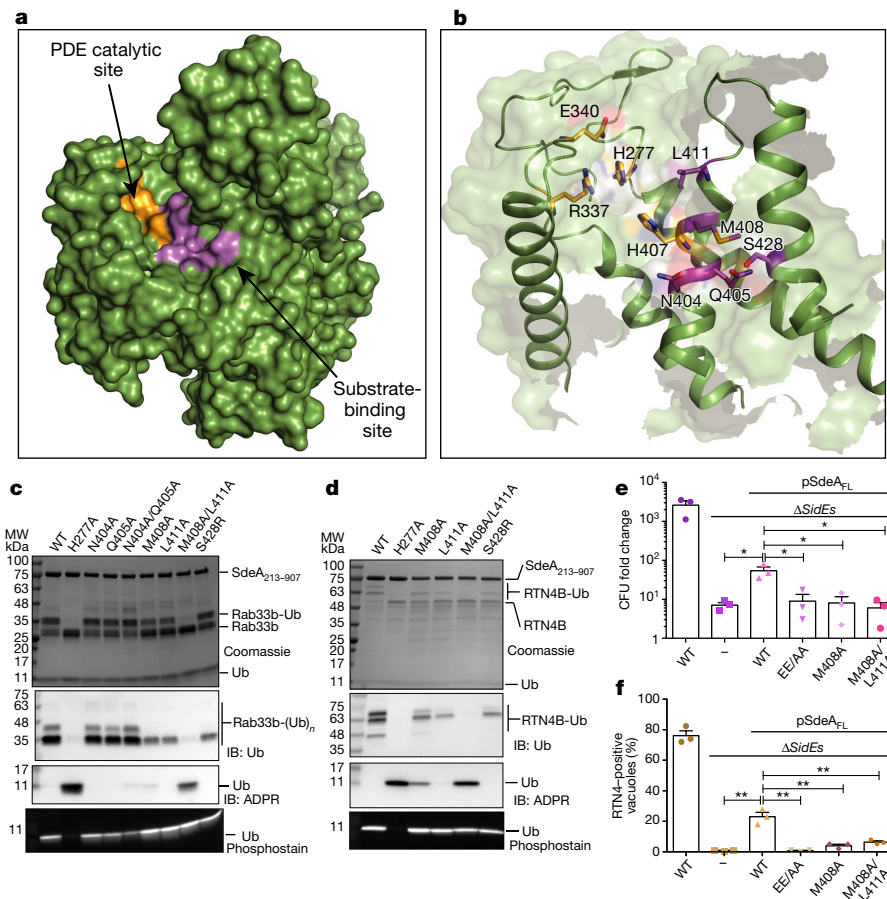


Fig. 4 | Substrate-binding site in SdeA PDE domain. **a**, SdeA PDE domain in surface representation with the catalytic site coloured orange. **b**, Amino acid residues in the PDE active site (orange) and the putative substrate-binding cleft (magenta). **c**, In vitro Rab33b ubiquitination assays with SdeA_{213–907} substrate-binding mutants. **d**, In vitro RTN4B ubiquitination assays with SdeA_{213–907} substrate-binding mutants. **e**, Fold change in colony forming units (CFU) in wild-type *L. pneumophila* and the Δ sidEs strain complemented with mutants defective in substrate ubiquitination. SdeA catalytic dead mutant EE/AA (E860/E862A) was used as a control ($n = 3$ biological replicates). Exact

P values: Δ sidEs versus pSdeA (SdeA plasmid) = 0.024; pSdeA versus pSdeA EE/AA = 0.032; pSdeA versus pSdeA M408A = 0.028; pSdeA versus pSdeA M408A/L411A = 0.023, as analysed by two-tailed *t*-test. **f**, Percentage of RTN4 positive vacuoles containing relevant *L. pneumophila* strains ($n = 3$ biological replicates). Exact *P* values: Δ sidEs versus pSdeA = 0.0015; pSdeA versus pSdeA EE/AA = 0.0015; pSdeA versus pSdeA M408A = 0.0034; pSdeA versus pSdeA M408A/L411A = 0.0051, as analysed by two-tailed *t*-test. Data shown as mean + s.e.m. **P* < 0.05, ***P* < 0.01 (**e**, **f**). Experiments shown in **c** and **d** were repeated independently twice with similar results. For gel source data, see Supplementary Fig. 1.

substrates for ubiquitination rather than the modification of ubiquitin is the central function of SdeA in acute bacterial pathogenicity.

The SdeA structure and the detailed biochemistry presented here give us a first-hand glimpse into the atomic details of the PR-ubiquitination catalysed by the SidE family of bacterial enzymes and enable us to pin down substrate ubiquitination as the pathogenic principle of SdeA. PR-ubiquitination by the PDE domain progresses through a transient intermediate in the form of SdeA H277–PR-Ub, which is subsequently attacked by the OH group of the target serine of the substrate for successful ubiquitin transfer. This suggests a double displacement mechanism for PDE catalysis in which the binding of PR-ubiquitination substrates at the PDE active site may first require release of AMP generated during the intermediate formation. This is consistent with the juxtaposed catalytic groove of PDE domain and the substrate binding cleft. Notably, the active sites of mART and PDE face opposite sides of the molecule (Fig. 1a, b), hinting that there may not be a direct transfer of ADP-ribosylated ubiquitin between the catalytic centres of the mART and PDE domains. However, potential dimerization of SdeA, as observed with purified proteins in solution (Extended Data Fig. 10), may enable these two catalytic sites to face each other *in trans*. Sequence analysis of SdeA substrates revealed that the target serine residues could occur in disordered regions in line with the limiting size of the substrate-binding cleft in the SdeA PDE domain (Fig. 4b). On the basis of the substrate recognition motif identified in

this study, we propose that SdeA could be a broad-specificity ligase that targets disordered serine residues in multiple substrates. Therefore, the specificity of SdeA-mediated ubiquitination during *Legionella* infection could be conferred by its recruitment to the endoplasmic reticulum, where currently identified *in vivo* substrates of SdeA reside^{2,9}. The dissection of PR-ubiquitination catalysis by SdeA presented here may also aid the future discovery of related mammalian enzymes.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0145-8>.

Received: 26 September 2017; Accepted: 23 April 2018;
Published online: 23 May 2018

- Hershko, A., Ciechanover, A. & Varshavsky, A. The ubiquitin system. *Nat. Med.* **6**, 1073–1081 (2000).
- Qiu, J. et al. Ubiquitination independent of E1 and E2 enzymes by bacterial effectors. *Nature* **533**, 120–124 (2016).
- Bhogaraju, S. et al. Phosphoribosylation of ubiquitin promotes serine ubiquitination and impairs conventional ubiquitination. *Cell* **167**, 1636–1649. e13 (2016).
- Wong, K., Kozlov, G., Zhang, Y. & Gehring, K. Structure of the *Legionella* effector, lpg1496, suggests a role in nucleotide metabolism. *J. Biol. Chem.* **290**, 24727–24737 (2015).

5. Ji, X. et al. Mechanism of allosteric activation of SAMHD1 by dGTP. *Nat. Struct. Mol. Biol.* **20**, 1304–1309 (2013).
6. Simon, N. C., Aktories, K. & Barbieri, J. T. Novel bacterial ADP-ribosylating toxins: structure and function. *Nat. Rev. Microbiol.* **12**, 599–611 (2014).
7. Barrio, J. R., Secrist, J. A. III & Leonard, N. J. A fluorescent analog of nicotinamide adenine dinucleotide. *Proc. Natl Acad. Sci. USA* **69**, 2039–2042 (1972).
8. Dong, Y. et al. Structural basis of ubiquitin modification by the *Legionella* effector SdeA. *Nature* <https://doi.org/10.1038/s41586-018-0146-7> (2018).
9. Kotewicz, K. M. et al. A single *Legionella* effector catalyzes a multistep ubiquitination pathway to rearrange tubular endoplasmic reticulum for replication. *Cell Host Microbe* **21**, 169–181 (2017).
10. Matte, A., Tari, L. W. & Delbaere, L. T. How do kinases transfer phosphoryl groups? *Structure* **6**, 413–419 (1998).
11. Kee, J.-M. & Muir, T. W. Chasing phosphohistidine, an elusive sibling in the phosphoamino acid family. *ACS Chem. Biol.* **7**, 44–51 (2012).
12. Fuhs, S. R. et al. Monoclonal 1- and 3-phosphohistidine antibodies: new tools to study histidine phosphorylation. *Cell* **162**, 198–210 (2015).
13. Lima, C. D., Klein, M. G. & Hendrickson, W. A. Structure-based analysis of catalysis and substrate definition in the HIT protein family. *Science* **278**, 286–290 (1997).
14. Akturk, A. et al. Mechanism of phosphoribosyl-ubiquitination mediated by a single *Legionella* effector. *Nature* <https://doi.org/10.1038/s41586-018-0147-6> (2018).
15. Thomsen, M. C. F. & Nielsen, M. Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Res.* **40**, W281–W287 (2012).
16. Nielsen, M. & Andreatta, M. NNAAlign: a platform to construct and evaluate artificial neural network models of receptor-ligand interactions. *Nucleic Acids Res.* **45**, W344–W349 (2017).

Acknowledgements We thank J. Vogel for the gift of anti-SidC serum, R. Prabu for help with total mass analysis of proteolysed SdeA, Ö. Yildiz for sharing synchrotron time, A. Chaikuad for initial construct design of SdeA, T. Hunter and C. Lima for advice on histidine intermediate protocols, T. Hanke for the PDE mechanism scheme and discussion, T. Colby and I. Matic for their help with identifying ubiquitination sites of RTN4B by LC-MS/MS and B. Schulman and D. Scott for crystallography advice. Swiss Light Source beamtime was part of proposal 20161958. We thank the staff of SLS for their assistance

in data collection as well as E. Veshkova, S. Rodriguez Gomez, S. Jelenic and F. Miljkovic for technical assistance; K. Koch, D. Höller, V. Dötsch and S. Knapp for comments on the paper; and D. Svergun's group at beamline P12, PETRA III, EMBL-DESY for SAXS data collection. This work was supported by iNEXT (PID:3515), the DFG-funded Collaborative Research Centre on Selective Autophagy (SFB 1177), the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 742720), the DFG-funded Cluster of Excellence "Macromolecular Complexes" (EXC115), the DFG-funded SPP 1580 program "Intracellular Compartments as Places of Pathogen-Host-Interactions" (I.D.) and the LOEWE program Ubiquitin Networks (Ub-Net) and the LOEWE Center for Gene and Cell Therapy Frankfurt (CGT), both funded by the State of Hesse/Germany. NIH-NIAID grant R01AI127465 supported Z.-Q.L. The work of S.B. is also funded by a Goethe University Nachwuchsforscher grant.

Reviewer information Nature thanks K. Gehring and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions S.K., S.B. and I.D. conceived the project. S.K. performed initial crystallization. S.K. and S.B. performed crystal optimization, structure solution, protein purification and biochemistry. J.B. contributed to crystallization and structure solution. F.B. performed mass spectrometry. D.S. performed SAXS and contributed to protein purification. Y.L. and P.G. performed experiments with SdeA_{FL}. N.G. performed bacterial infection experiments. S.K., S.B., Z.-Q.L. and I.D. analysed the data. S.K., S.B. and I.D. wrote the manuscript. I.D. supervised the project.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0145-8>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0145-8>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to I.D.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Expression and purification. Expression and purification of SdeA and Rab33b have previously been described². In brief, T7 express cells were transformed with wild-type SdeA (UniProt accession code: Q5ZTK4) or mutant constructs cloned in a modified pET21a vector with a C-terminal CPD (cysteine protease domain)–His tag. Rab33b was cloned into a pET21a vector with a C-terminal His tag. For selenomethionine-labelled SdeA_{213–907}, the plasmid was transformed into B834 competent cells and was expressed using selenomethionine-containing minimal medium (Molecular Dimensions). T7 express transformed cells were grown in LB medium at 37 °C to an optical density (OD) of 0.6–0.8, induced with 0.5 mM IPTG (isopropyl β-D-1-thiogalactopyranoside), grown overnight at 18 °C and collected. The cell pellet was resuspended in a buffer containing 50 mM Tris-HCl pH 7.5, 300 mM NaCl, 10% (v/v) glycerol, 1 mM PMSE, DNase and protease inhibitor cocktail tablets (Roche). The cells were lysed by sonication. Clarified supernatant (12,000 r.p.m., 4 °C, 40 min) was incubated for 1 h with pre-equilibrated Talon beads before being washed three times with 50 mM Tris-HCl pH 7.5, 300 mM NaCl, 10% (v/v) glycerol. For RTN4B purification, the clarified supernatant was spun again (40,000 r.p.m., 4 °C, 90 min) to collect membrane fractions. The membranes were homogenized and solubilized in the presence of 1% n-dodecyl β-D-maltoside (DDM). For the rest of the RTN4B purification, 0.05% (w/v) DDM was maintained in the buffer. For Rab33b purification, the protein was eluted using 200 mM imidazole after binding to Talon beads. The CPD–His tag of SdeA and RTN4B was cleaved off the protein while it was still bound to Talon beads using 100 μM phytic acid. For biochemical assays, the buffer was exchanged to a final buffer of 10 mM HEPES pH 7.5, 150 mM NaCl and 1 mM TCEP. For crystallization, SdeA_{213–907} was loaded onto a Q-sepharose column after CPD tag cleavage. SdeA_{213–907} eluted in flowthrough, while most of the impurities were bound to the column. The fraction containing the protein was then concentrated before injecting into a Superdex 75 16/60 size-exclusion chromatography column pre-equilibrated with 10 mM HEPES pH 7.5, 150 mM NaCl, 1 mM TCEP. The protein eluted in a single peak and fractions were pooled together and concentrated to 25 mg/ml before setting up crystallization screens.

Limited proteolysis. One milligram of SdeA_{193–998} was incubated with 50 μg GluC in 20 mM HEPES pH 7.5, 50 mM NaCl and 10 mM MgSO₄ for 1 h on ice. This was followed by size-exclusion chromatography of the reaction on a Superdex 75 16/60 column equilibrated with 10 mM HEPES pH 7.5, 150 mM NaCl. The fractions were pooled and protein matrix assisted laser desorption/ionization (MALDI) mass spectrometry was performed. The protein fragment of interest was then identified based on the exact mass using the ExPASy server tool Findpept. Various constructs of SdeA were also similarly proteolysed by GluC and analysed by Coomassie stained SDS gel.

Crystallization. The purified protein was clarified by centrifugation (10,000 r.p.m., 10 min, 4 °C) before we set up crystallization plates. Sitting drop and hanging drop sparse matrix screens in 96-well format were set up with 125 nl protein and 125 nl precipitant solution. The protein crystallized in 100 mM Bis Tris propane pH 7.0–8.0, 0.1–0.2 M sodium citrate tribasic dihydrate, 20–30% (w/v) PEG 3350 at 20 °C. The morphology of the thin plate-like crystals was improved using 0.5–1% (v/v) ethylene glycol and 50–200 mM non-detergent sulfobetaine-201 (NDSB-201) as additives. The crystals were cryoprotected using mother liquor supplemented with 15% (v/v) glycerol and 10% (v/v) ethylene glycol and flash frozen in liquid nitrogen. SeMet-labelled SdeA crystallized in similar conditions as wild-type SdeA.

Data collection, data processing and structure solution. Both SeMet and native diffraction data were collected at the PXIII beamline of the Swiss Light Source, Villigen. Native data were collected at wavelength 1.00003 Å. The data were processed using XDS¹⁷. SAD (single anomalous dispersion) data from SeMet crystals were collected at wavelength 0.97927 Å and processed using XDS, and the phases were calculated using Phenix autosol¹⁸. Using these initial phases, buccaneer was used to build most of the PDE domain and segments of mART and AHL that were well-ordered¹⁹. After manual correction of the output model of buccaneer, MR-SAD (molecular replacement–single-wavelength anomalous dispersion) was performed using PHENIX autosol to further improve the phases. Further building was done manually in Coot using SeMet positions as a guide²⁰. Using the model obtained from experimental phases, molecular replacement was performed for the native data of SdeA extending to 2.8 Å. Further iterative cycles of manual building and refinement were performed using Coot and refinement programs phenix refine and Buster (Global phasing)^{18,21}. Ramachandran statistics for the refined SdeA core structure are favoured: 93.6%, allowed: 6.4%, outliers: none.

In vitro ubiquitination assays. SdeA ubiquitination experiments were done as described³. In brief, 2.5 μg purified untagged ubiquitin and 2 μg Rab33b were incubated with 1 μg SdeA (FL, 213–907 wild-type and variants and SdeA PDE domain) at 37 °C for 30 min in the presence or absence of 200 μM NAD⁺ in a buffer containing 50 mM Tris-HCl pH 7.5, 50 mM NaCl in a final reaction volume of 30 μl. For reactions involving the SdeA PDE domain, purified ADP ribosylated ubiquitin (2.5 μg) was used instead of NAD⁺. ADP ribosylated ubiquitin

was generated using the SdeA H277A mutant and purified using size-exclusion chromatography. The reaction mixture was subjected to SDS–PAGE followed by Coomassie staining. Alternatively, the reaction mixture was subjected to SDS–PAGE followed by western blotting using ubiquitin antibody and pan-ADP ribose antibody (Millipore) or to phosphostaining to identify PR-UB (Pro-Q diamond phosphostaining protocol, Thermo Fisher). For ubiquitination assays using FL GFP-tagged SdeA constructs, we transfected 2 μg GFP-tagged SdeA constructs into HEK293T cells cultured in 6-well plates. HEK293T cells were obtained from ATCC (ATCC CRL-3216) and authenticated using STR DNA profiling. All the cell lines used tested negative for mycoplasma. Cells were collected after 24 h and lysed in 150 μl lysis buffer (50 mM Tris-HCl pH 7.4, 150 mM NaCl, 1% Triton X-100, 10% glycerol, protease inhibitors cocktail and 1 mM PMSF). Using 15 μl GFP-trap beads, we purified the GFP-tagged SdeA proteins from the clarified lysate. After extensive washing of the beads, an in vitro ubiquitination reaction was set up in 30 μl reaction buffer (50 mM Tris-HCl, 50 mM NaCl, pH 7.4) with 3 μg purified Rab33b, 2.5 μg ubiquitin and 0.2 mM NAD⁺. The reaction was performed at 37 °C in a thermomixer while shaking. After 30 min, the reaction was stopped by adding SDS loading buffer and the samples were analysed using Coomassie staining of SDS gel and western blotting. The reaction mixture was analysed similarly to the assays using bacterially purified SdeA_{213–907} except for the use of complementary ubiquitin antibodies (CS-Ub and abcam-Ub) to monitor ubiquitin modification³. Where indicated, ADPR-Ub was used as co-factor in reactions instead of Ub and NAD. Peptide ubiquitinations were carried out with 0.5 mM of each peptide as substrate with SdeA_{FL}. All experiments were repeated at least twice.

ε-NAD⁺ hydrolysis. For measuring the ubiquitin ADP-ribosylation kinetics of SdeA_{213–907} and its mutants, we used an ε-NAD hydrolysis assay⁷. Five micrograms of wild-type or mutant SdeA was incubated in a buffer containing 50 mM Tris pH 7.5, 50 mM NaCl with 100 μg ubiquitin in 100 μl reaction buffer. ε-NAD was added to final concentration of 1 mM to start the reaction. Fluorescence of ε-adenine (excitation wavelength: 300 nm, emission wavelength: 410 nm) was monitored using a plate reader at 25 °C at every 1-min interval. Constructs of the C-terminal region (909–1499 and 909–1233) were added in 1.5 molar excess of SdeA_{213–907} to test their effect on activity. All experiments were repeated at least twice.

Histidine intermediate analysis. This analysis was performed as described previously with slight modifications¹². In brief, 4 μg wild-type or mutant SdeA_{213–907} was incubated with 10 μg ubiquitin or labelled ubiquitin and 2 mM NAD⁺ for 5 min at 37 °C. The reaction was stopped by transferring the contents onto ice and adding 5× SDS loading buffer (pH 8.8). Gel electrophoresis and transfer were conducted at 4 °C and analysed by Coomassie staining, fluorescence scanning or western blotting. All experiments were repeated at least twice.

Small-angle X-ray scattering. SdeA_{213–907} was purified in 10 mM HEPES pH 7.5, 150 mM NaCl, 1 mM TCEP by size-exclusion chromatography (Superdex 200 increase). Each fraction was collected and concentrated with centrifugal concentration devices (50-kDa cut-off, Supplementary Table 2, Supplementary Information). The scattering profile of flow-through buffer from the size-exclusion column was recorded as reference buffer scattering. SAXS data were collected at beamline P12, EMBL-DESY (Supplementary Table 2, Supplementary Information). Primary data analysis was performed using PRIMUS from the ATSAS package²². Bead modelling was conducted with DAMMIF²³ and DAMMIN²⁴ from the ATSAS package. SAXS curves from atomic model were generated and fitted to experimental data using CRYSOLO²⁵ and SUPCOMB²⁶ from ATSAS 2.8.3 package.

Mass spectrometry. For analysis of the histidine intermediate, an in vitro ubiquitination reaction of SdeA_{213–907}(H407N) was stopped after 5 min by denaturation on ice in 5.3 M urea pH 8.8 for 10 min. The sample was diluted to 2 M urea with 50 mM ABC pH 8.8 and was loaded onto a 30-kDa filter (Amicon Ultra, 0.5 ml, Merck). (Ub-)SdeA was trypsinized according to an adapted FASP-protocol as previously described³. In brief, the proteins were washed four times with 100 μl ABC and after adding Trypsin Gold (Promega) in an enzyme:protein ratio of 1:2, tryptic digestion was performed for 20 min at 22 °C. Tryptic peptides in 50 mM ammonium bicarbonate pH 8.8 were loaded onto a 15-cm self-packed C18 column and separated with a short gradient (12 min from 10–38% buffer B (80% acetonitrile, 0.1% formic acid)) on an easy nLC2 system and injected online into a Q Exactive HF mass-spectrometer. Targeted MS2 scans were used to specifically fragment the bridged intermediate with different collision energies. For partial and specific fragmentation of the phosphoramidate bond, a normalized collision energy (NCE) of 15 was applied, and for fragmentation of the peptide backbone NCE 30 was used. Spectra were annotated manually and StavroX 3.6 was used for additional identification of the peptide backbone fragments²⁷. Identification of phosphoribose-bridged ubiquitination sites in RTN4B was done as described before by HCD and targeted ETD fragmentation after a modified FASP digest³.

L. pneumophila strains and host infections. *L. pneumophila* strains used in this study were derivatives of the Philadelphia 1 strain Lp02¹ and were grown and maintained on CYE (charcoal-yeast extract) plates or in *N*-(2-acetamido)-2-aminoethane (ACES) buffered yeast extract (AYE) broth as previously described²⁸.

The *sde* family in-frame deletion strain and complementation strains have been described previously². *sdeAM408A* and *sdeAM408A_L411A* mutant genes were cloned into pZL507³ for complementation. Raw264.7 cells were cultured in RPMI 1640 medium supplemented with 10% FBS. *D. discoideum* AX4 cells were cultured in HL-5 medium and maintained in MB medium for infection as described². Infection experiments were performed in triplicate.

Infection. *L. pneumophila* strains were grown to the post-exponential phase ($OD_{600} = 3.0\text{--}3.6$) in AYE broth. Complementation strains were induced with 0.2 mM IPTG for 4 h at 37 °C before infection. Raw 264.7 cells were infected with *L. pneumophila* strains at an MOI of 10 for 2 h to detect the translocation of SdeA and its mutants. Raw cells were collected and lysed with 0.2% saponin on ice for 30 min. Cleared cell lysates were resolved by SDS–PAGE, followed by immunoblotting with antibodies specific for SdeA and tubulin. Total *L. pneumophila* proteins were resolved by SDS–PAGE to evaluate the expression of SdeA by immunoblotting with SdeA-specific antibodies, and isocitrate dehydrogenase (ICDH) was probed as a loading control with previously described antibodies²⁹. For intracellular growth in *D. discoideum*, infection was performed at an MOI of 0.1 and the total bacterial counts were determined at 24-h intervals as described³⁰. The enrichment of RTN4 by bacterial phagosomes was assessed by immunostaining in primary mouse macrophages infected with the relevant *L. pneumophila* strains for 2 h at an MOI of 1.0. Primary mouse macrophages were obtained from A/J mouse (female, 6 weeks, Jackson lab cat#000646). No randomization of mice and blinding was necessary as mice were used only to collect primary bone marrow macrophages for *Legionella* infection experiments. Immunostaining with anti-RTN4 (Lsbio cat#LS-B6516-50) (1:500) was performed as described⁹. Infection experiments were performed in triplicate.

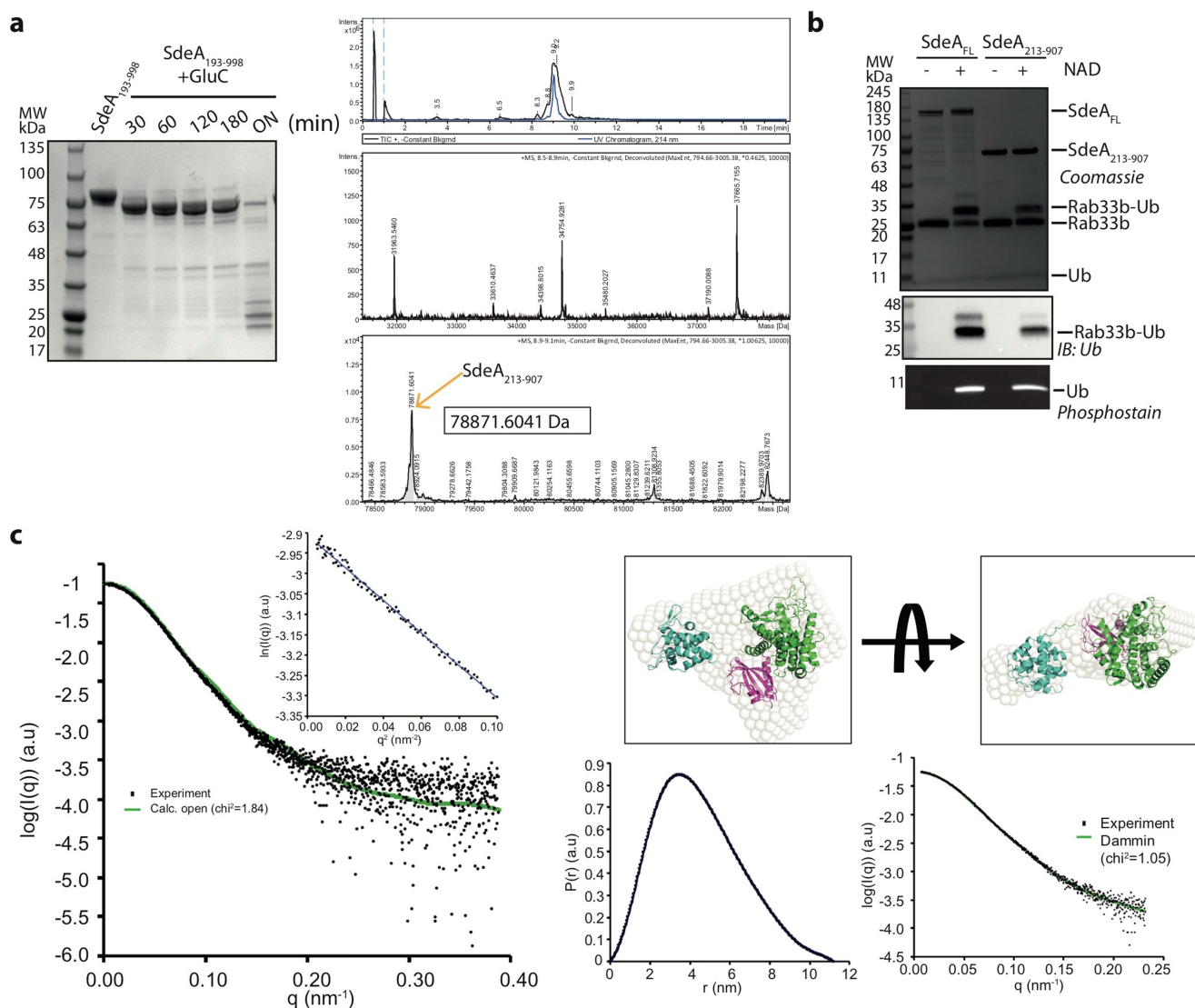
Antibodies and immunoblotting. For immunoblotting, samples resolved by SDS–PAGE were transferred onto 0.2- μ m nitrocellulose membranes (Pall Life Sciences cat#66485). Membranes were blocked with 5% non-fat milk and incubated with the appropriate primary antibodies: anti-SdeA2, 1:10,000, anti-ICDH3, 1:10,000, anti-tubulin (DSHB, E7) 1:10,000. Membranes were then incubated with an appropriate IRDye infrared secondary antibody (dilution: 1:20,000) and scanned using an Odyssey infrared imaging system (Li-Cor's Biosciences).

Ethical compliance. Animal protocols used in the study were approved by Purdue Animal Care and Use Committee. We complied with all the relevant ethical regulations. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

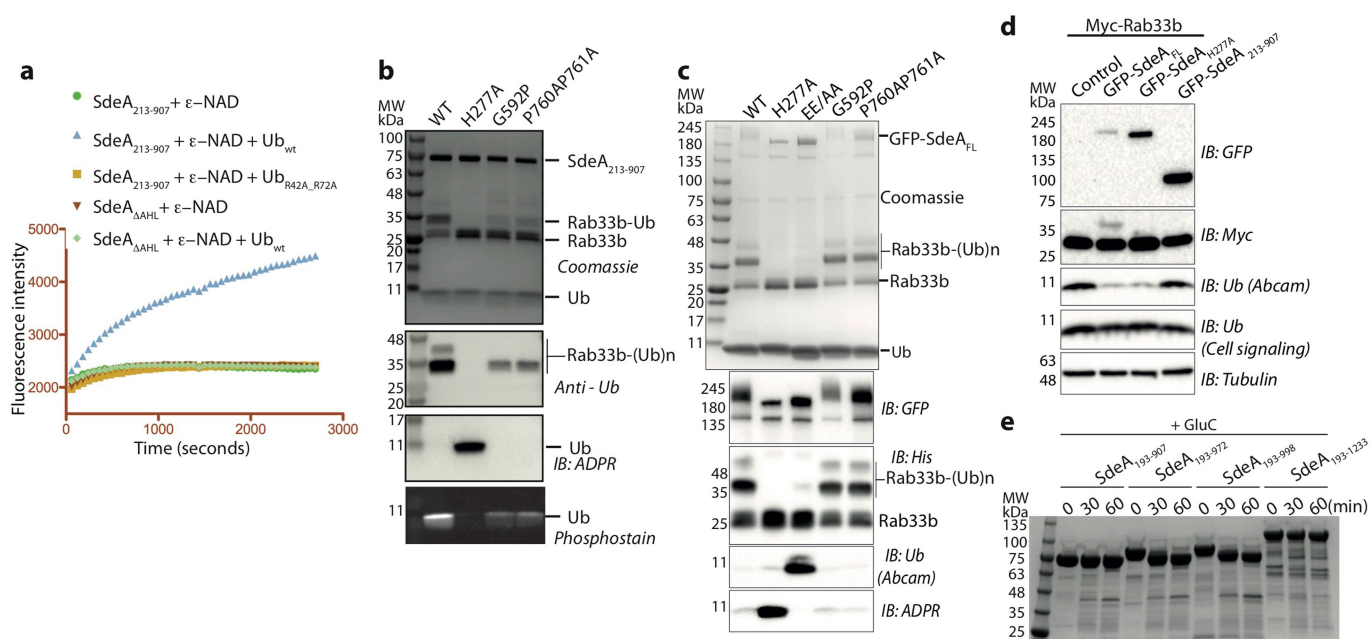
Data availability. Structure coordinates are available from the Protein Data Bank under accession code 6G0C. Small-angle X-ray scattering data and models are available from SASBDB (<https://www.sasbdb.org/>) under accession number SASDD65. Full gel source data can be found in Supplementary Fig. 1. The data that support the findings of this study are available from the corresponding author upon request.

- Kabsch, W. XDS. *Acta Crystallogr. D* **66**, 125–132 (2010).
- Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
- Cowan, K. The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr. D* **62**, 1002–1011 (2006).
- Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
- Bricogne, G. et al. BUSTER version 2.10.2 (United Kingdom Global Phasing Ltd., Cambridge, 2017).
- Franke, D. et al. ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J. Appl. Crystallogr.* **50**, 1212–1225 (2017).
- Franke, D. & Svergun, D. I. DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. *J. Appl. Crystallogr.* **42**, 342–346 (2009).
- Svergun, D. I. Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophys. J.* **76**, 2879–2886 (1999).
- Svergun, D., Barberato, C. & Koch, M. H. J. IUCr. CRYSOLO – a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.* **28**, 768–773 (1995).
- Kozin, M. B. & Svergun, D. I. IUCr. Automated matching of high- and low-resolution structural models. *J. Appl. Crystallogr.* **34**, 33–41 (2001).
- Götze, M. et al. StavroX—a software for analyzing crosslinked products in protein interaction studies. *J. Am. Soc. Mass Spectrom.* **23**, 76–87 (2012).
- Berger, K. H. & Isberg, R. R. Two distinct defects in intracellular growth complemented by a single genetic locus in *Legionella pneumophila*. *Mol. Microbiol.* **7**, 7–19 (1993).
- Xu, L. et al. Inhibition of host vacuolar H⁺-ATPase activity by a *Legionella pneumophila* effector. *PLoS Pathog.* **6**, e1000822 (2010).
- Luo, Z.-Q. & Isberg, R. R. Multiple substrates of the *Legionella pneumophila* Dot/Icm system identified by interbacterial protein transfer. *Proc. Natl Acad. Sci. USA* **101**, 841–846 (2004).



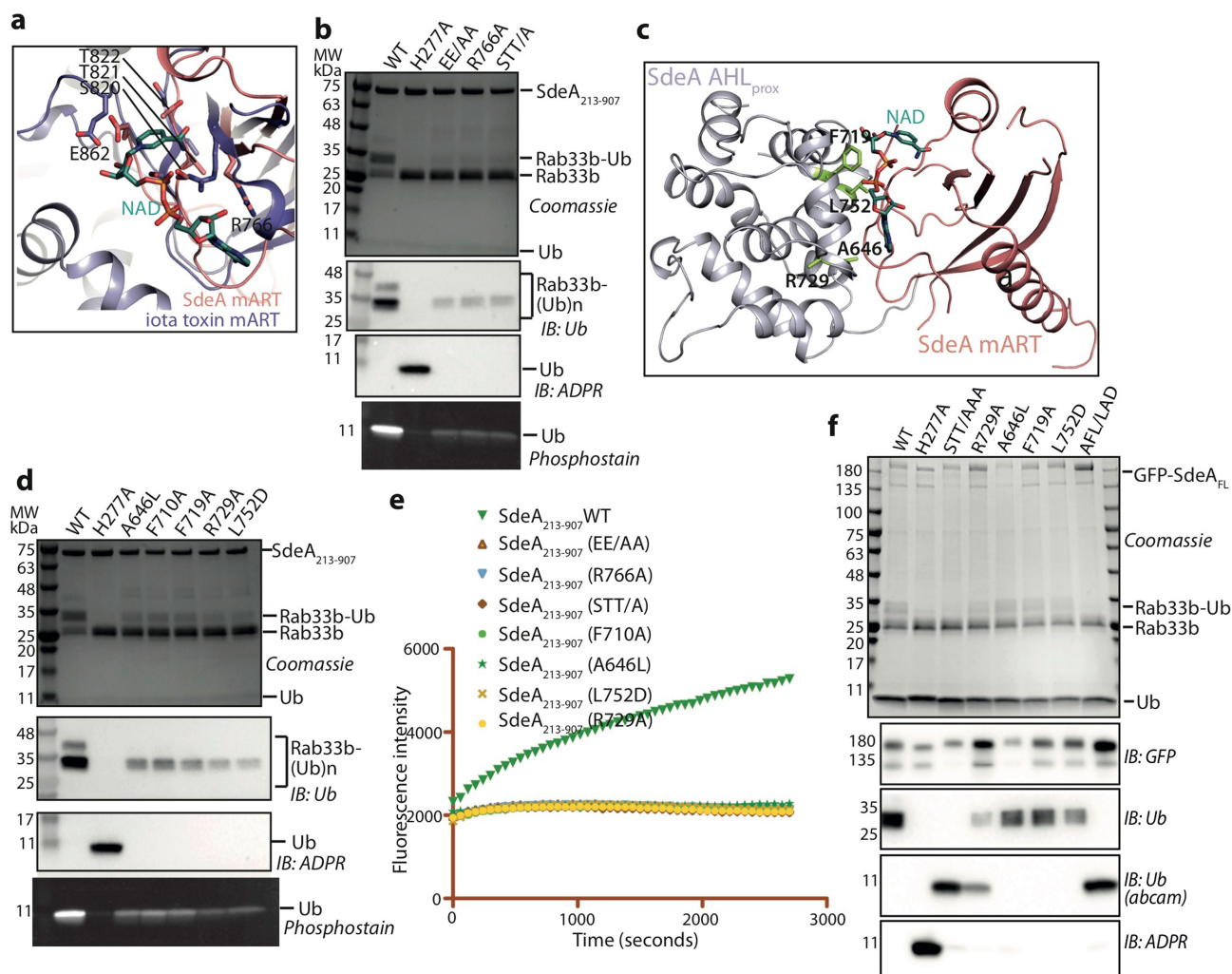
Extended Data Fig. 1 | Catalytic core—SdeA₂₁₃₋₉₀₇. **a**, Limited proteolysis of SdeA fragment 193–998 and subsequent analysis of the fragments by Coomassie-stained SDS gel and total mass analysis by mass spectrometry. **b**, In vitro ubiquitination of Rab33b by SdeA_{FL} and SdeA₂₁₃₋₉₀₇. **c**, Left, scattering profile of SdeA₂₁₃₋₉₀₇ with calculated scattering curve from crystal structure. Gunier region is shown in inset. Top right, ab initio bead

model from DAMMIN superimposed with crystal structure and shown in two orientations. Bottom right, pair distance distribution plot and DAMMIN model fitting results. Experiments were repeated independently twice with similar results (**a**, **b**). For gel source data, see Supplementary Fig. 1.



Extended Data Fig. 2 | Role of AHL in SdeA. **a**, ε-NAD⁺ hydrolysis assay in the presence of SdeA (SdeA₂₁₃₋₉₀₇ or SdeA₂₁₃₋₉₀₇ΔAHL) and ubiquitin (wild-type or R42A_R72A). **b**, **c**, In vitro ubiquitination assay with mutations in loops connecting AHL to PDE and the MART catalytic core in SdeA₂₁₃₋₉₀₇ (**b**) and in SdeA_{FL} (**c**). **d**, Substrate ubiquitination and ubiquitin modification by SdeA_{FL} and SdeA₂₁₃₋₉₀₇ in HEK293T cells.

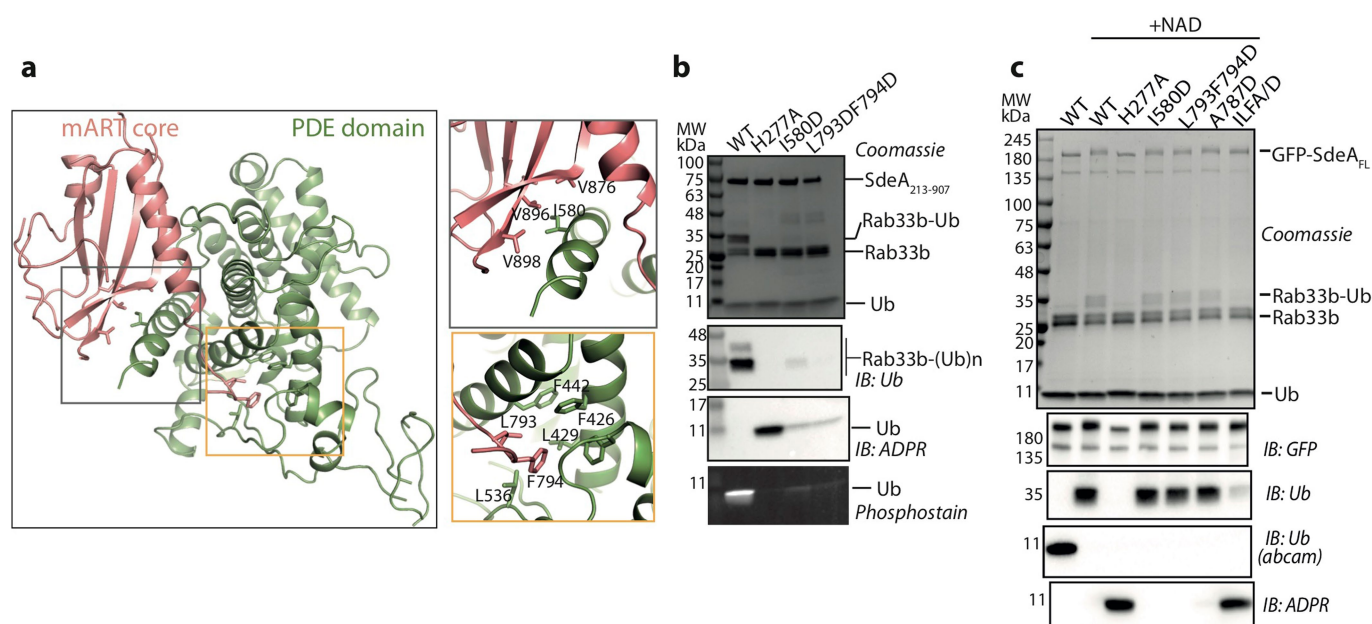
Abcam Ub and Cell Signalling Ub antibodies were used to monitor the levels of unmodified ubiquitin and total ubiquitin, respectively. **e**, Limited proteolysis analysis of various SdeA constructs. All experiments were repeated independently twice with similar results. For gel source data, see Supplementary Fig. 1.



Extended Data Fig. 3 | Characterization of the mART domain.

a, Superimposition of the mART core of SdeA with that of NAD⁺ bound structure of iota toxin (PDB: 4H0Y) from *Clostridium perfringens*, which ADP-ribosylates actin of host cells. Residues in SdeA that are predicted to be important for NAD⁺ binding and hydrolysis are labelled. **b**, In vitro ubiquitination assay with NAD⁺ binding site mutants in the mART core of SdeA₂₁₃₋₉₀₇. **c**, Residues at the interface between the mART core and AHL in proximal conformation (AHL_{prox}). **d**, In vitro ubiquitination assays

with the mutants of SdeA₂₁₃₋₉₀₇ mART core-AHL_{prox} interface residues indicated in c. **e**, Comparison of ϵ -NAD⁺ hydrolysis by SdeA₂₁₃₋₉₀₇ and NAD⁺ binding site mutants and mutants disrupting the predicted mART core-AHL_{prox} interaction. **f**, In vitro ubiquitination assays with the mutants of SdeA_{FL} mART core-AHL_{prox} interface residues indicated in c. Experiments were repeated independently twice with similar results. For gel source data, see Supplementary Fig. 1.

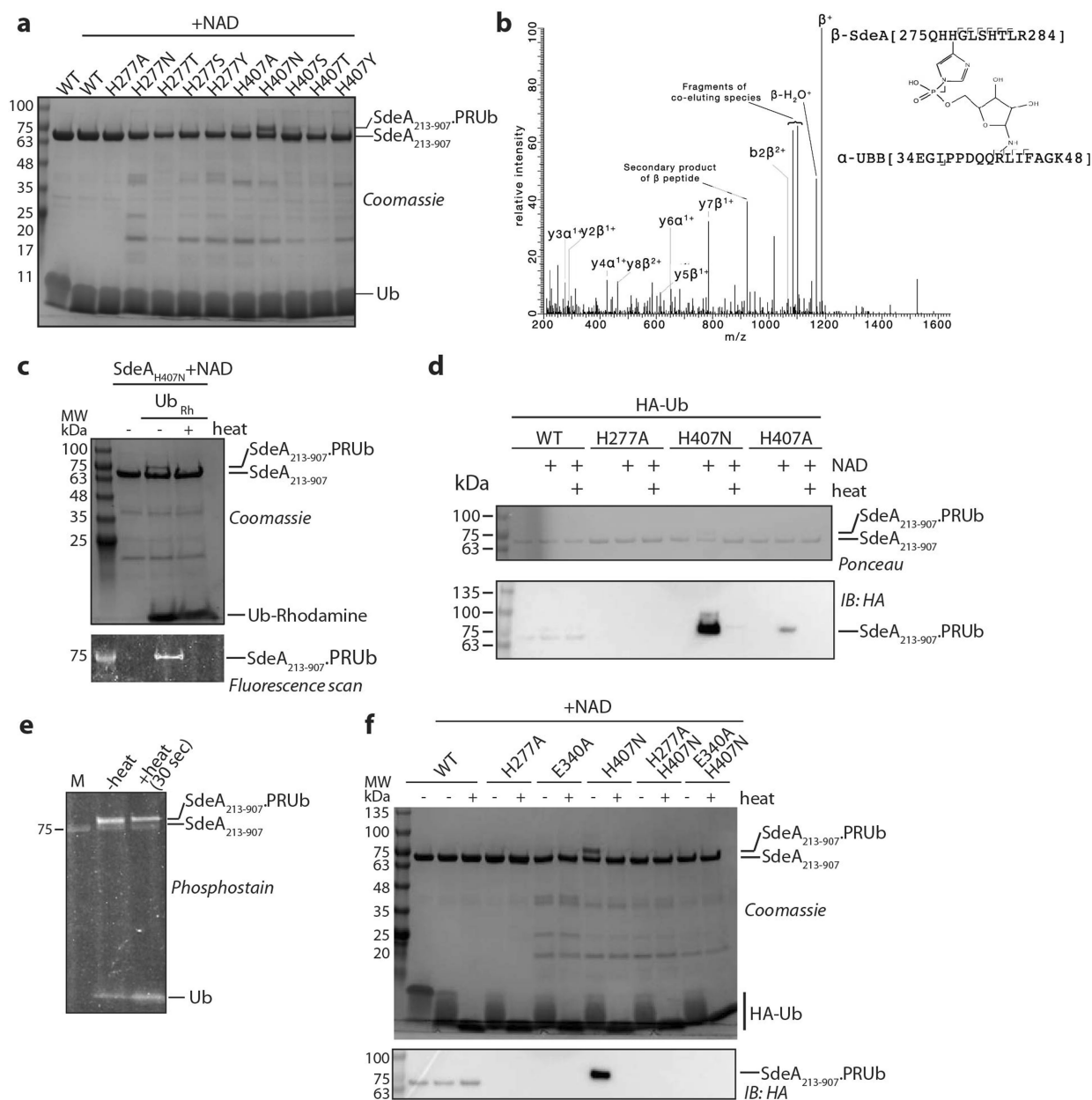


Extended Data Fig. 4 | Interaction between PDE and mART core.

a, Details of interaction between SdeA PDE and mART core. Important residues mediating the interaction are indicated in the insets.

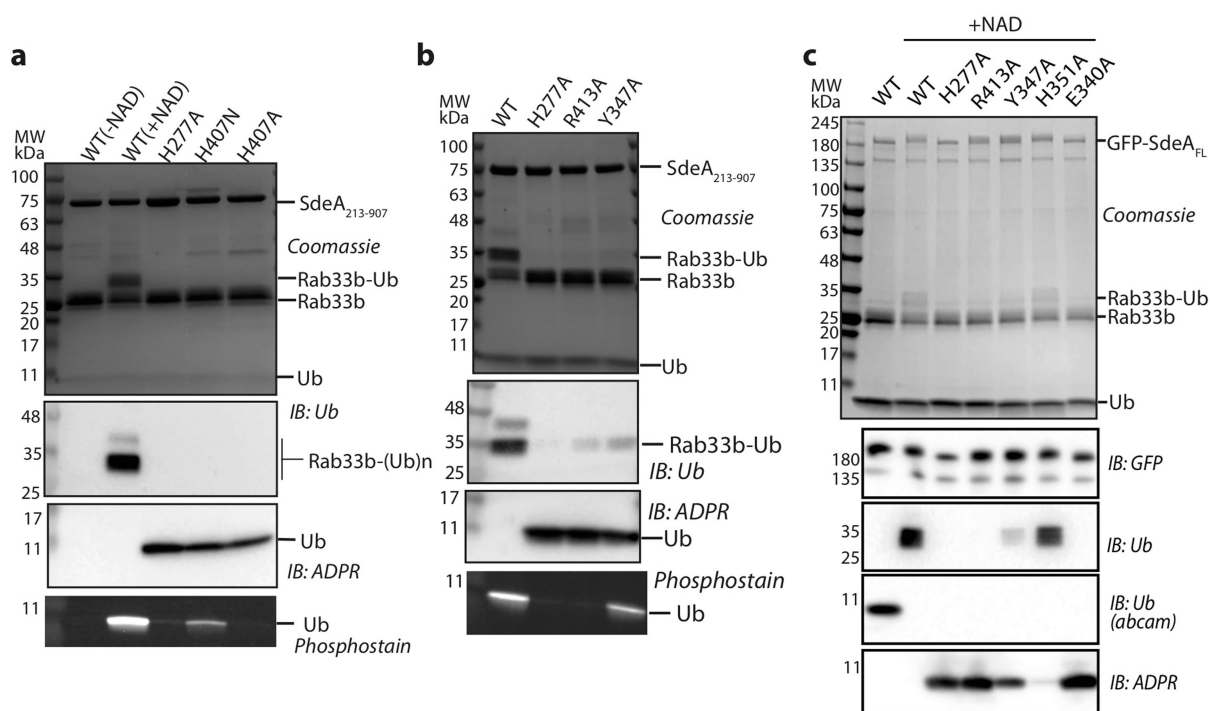
b, c, Testing in vitro substrate ubiquitination and ubiquitin modification

by PDE–mART core interaction mutants in SdeA₂₁₃₋₉₀₇ (**b**) and SdeA_{FL} (**c**). Experiments were repeated twice independently with similar results. For gel source data, see Supplementary Fig. 1.



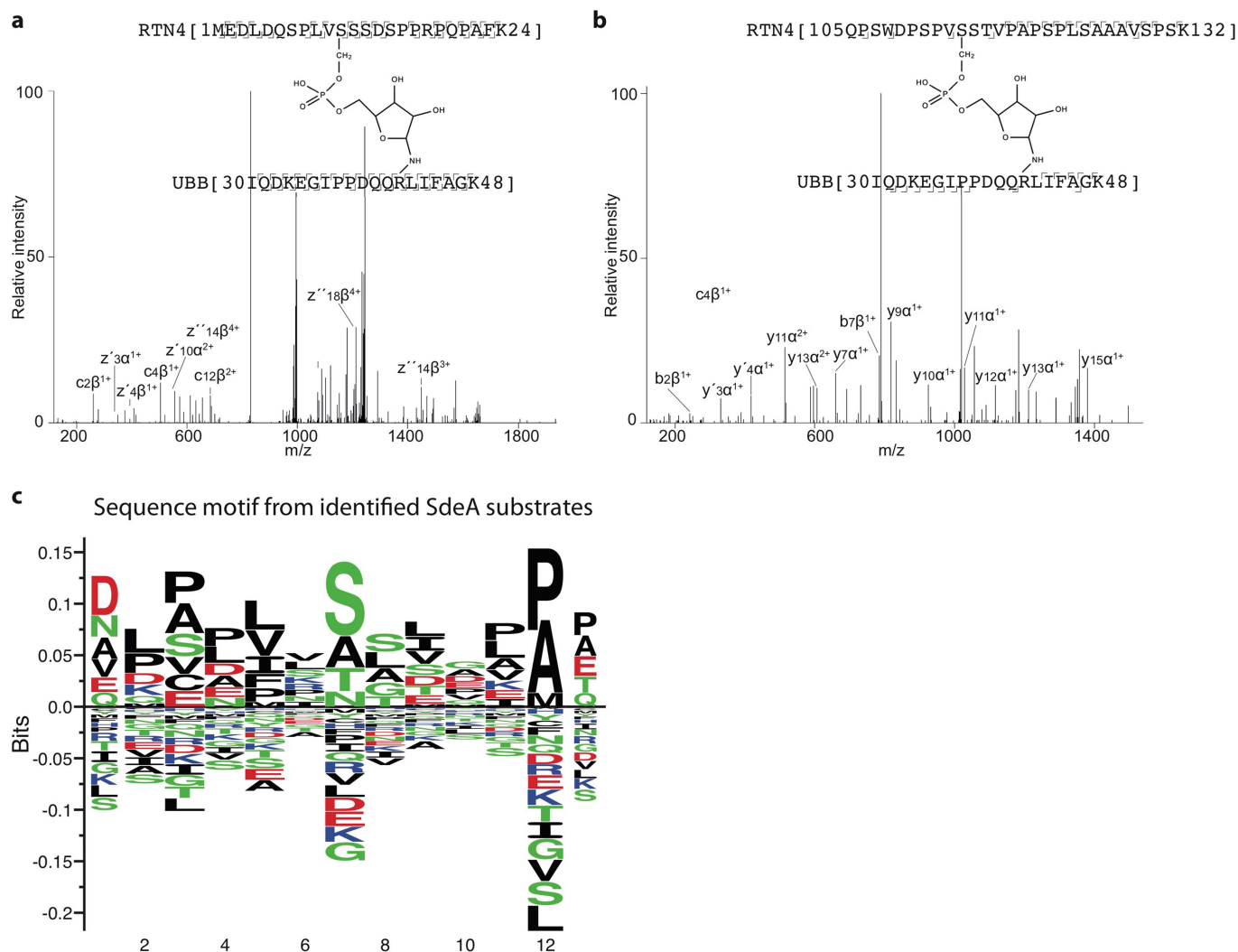
Extended Data Fig. 5 | Histidine intermediate in SdeA catalysis. **a**, In vitro ubiquitination reactions by various SdeA₂₁₃₋₉₀₇ PDE site histidine mutants probed by Coomassie-stained SDS-PAGE. **b**, High-energy HCD fragmentation was used to generate fragments of the peptide backbone. We could identify multiple fragments of SdeA₂₇₅₋₂₈₄ and Ub₃₄₋₄₈, to further validate the identity of the bridged active site. **c**, In vitro ubiquitination reaction by SdeA₂₁₃₋₉₀₇(H407N) mutant using rhodamine-labelled

ubiquitin. **d**, In vitro ubiquitination reaction using HA-tagged ubiquitin. **e**, In vitro ubiquitination reaction by SdeA₂₁₃₋₉₀₇(H407N) without and with heating probed by phosphostain. **f**, In vitro ubiquitination reaction using HA-ubiquitin by various PDE mutants. These experiments were repeated independently twice with similar results. For gel source data, see Supplementary Fig. 1.



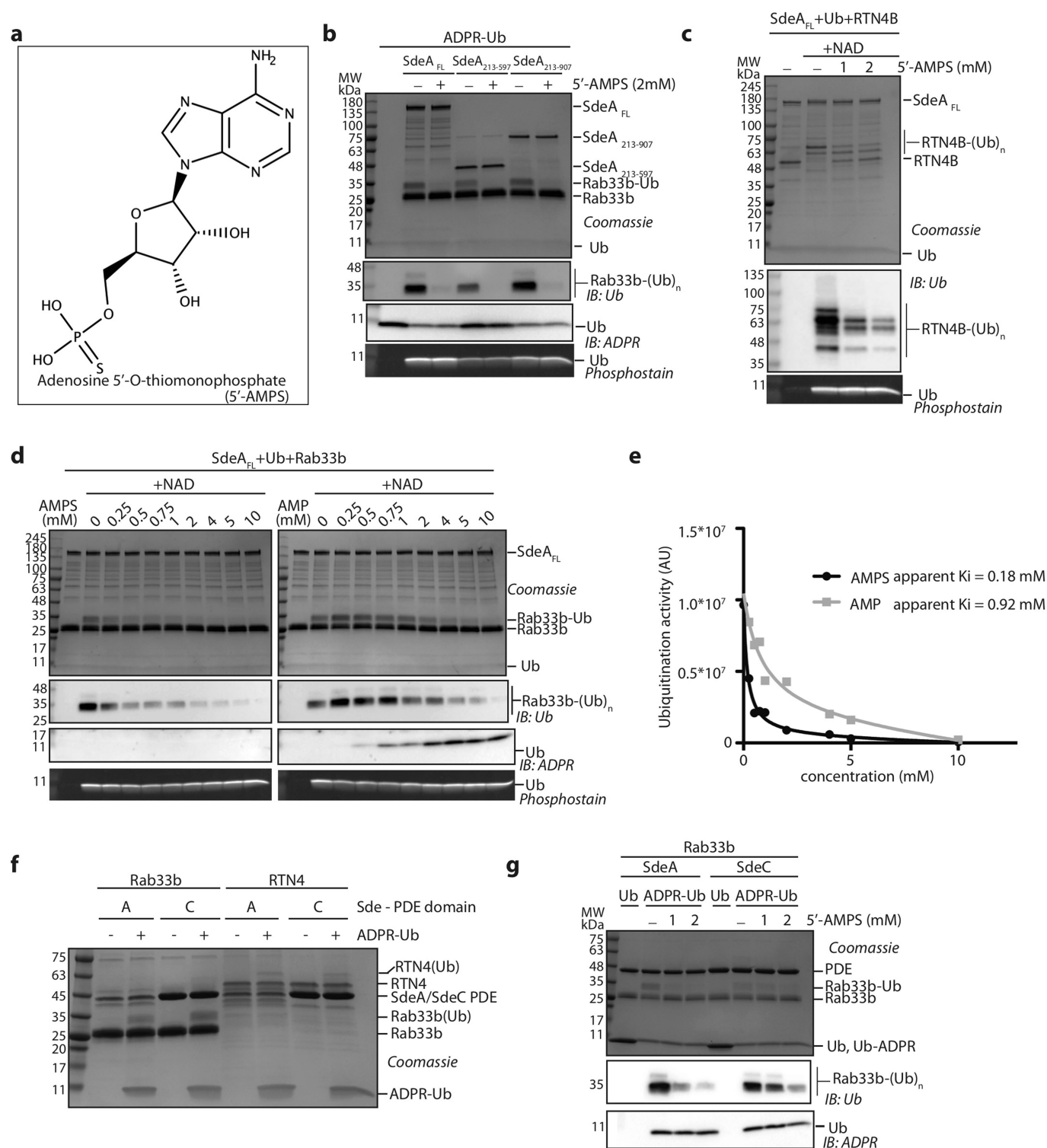
Extended Data Fig. 6 | PDE domain catalytic site. **a**, In vitro Rab33b ubiquitination by SdeA PDE mutants. **b**, In vitro ubiquitination assays with PDE catalytic site mutants. **c**, In vitro Rab33b ubiquitination assays

with GFP-SdeA_{FL} PDE catalytic site mutants purified from HEK293T cells. These experiments were repeated independently twice with similar results. For gel source data, see Supplementary Fig. 1.



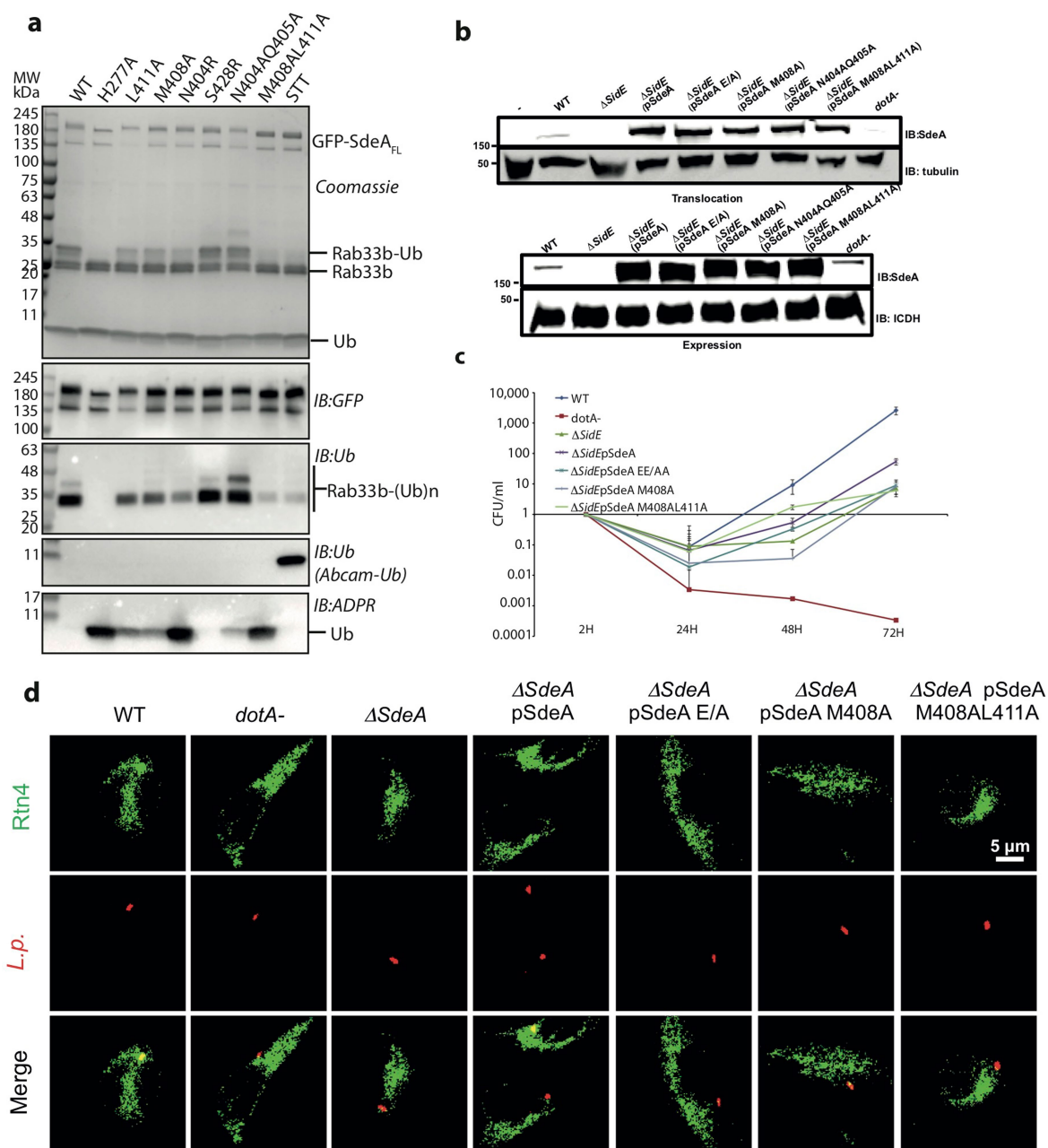
Extended Data Fig. 7 | Substrate specificity of SdeA. **a, b,** Fragmentation spectra of the bridged peptide indicating RTN4B ubiquitination sites.

These experiments were done once. **c,** Sequence motif of target serine sequences of SdeA as computed by Seq2Logo.



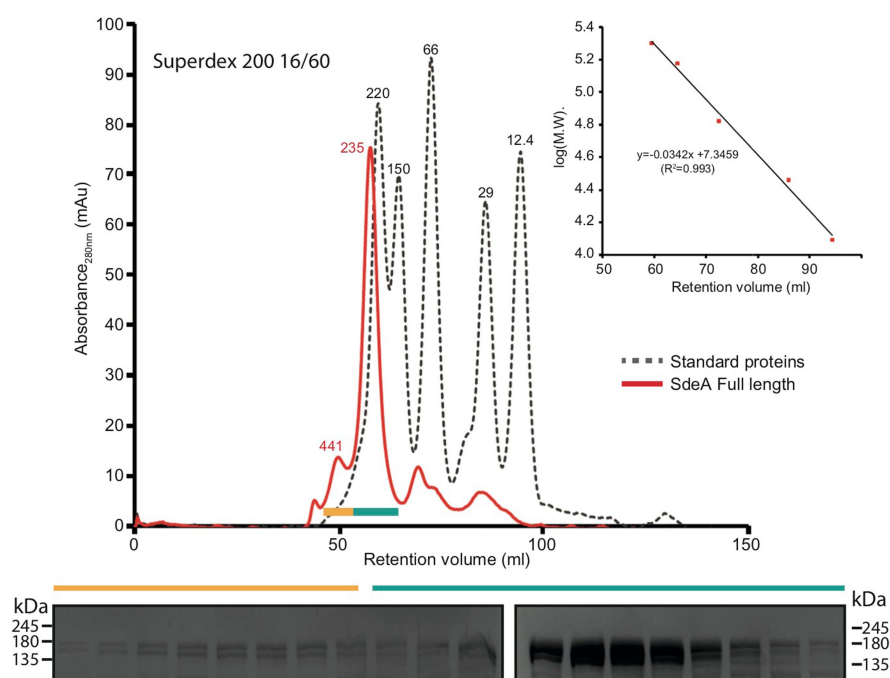
Extended Data Fig. 8 | Chemical inhibition of SdeA. **a**, Chemical structure of adenosine-5'-thio-monophosphate (5'-AMPS). **b**, 5'-AMPS-mediated inhibition of Rab33b PR-ubiquitination by SdeA_{FL}, SdeA₂₁₃₋₅₉₇ and SdeA₂₁₃₋₉₀₇ in the presence of ADPR-Ub. **c**, 5'-AMPS-mediated inhibition of RTN4B PR-ubiquitination by SdeA_{FL}. **d**, In vitro ubiquitination by SdeA_{FL} in the presence of increasing concentrations

of 5'-AMPS and AMP. **e**, Apparent inhibition constants (K_i) of AMP and 5'-AMPS against SdeA_{FL} calculated from quantification of substrate ubiquitination shown in **d**. **f**, PDE domain of SdeC ubiquitinates Rab33b and RTN4B. **g**, Effect of 5'-AMPS on ubiquitination by SdeC PDE. These experiments were done twice independently with similar results. For gel source data, see Supplementary Fig. 1.



Extended Data Fig. 9 | Effect of SdeA substrate-binding mutations in vivo. **a**, In vitro Rab33b ubiquitination assays with GFP-SdeA_{FL} substrate-binding mutants purified from HEK293T cells. **b**, Expression and translocation of SdeA using wild-type and various mutant strains of *Legionella*. **c**, CFU fold change monitored in wild-type *Legionella* and Δ *sidEs* strain complemented with substrate ubiquitination defective

mutant plasmids ($n = 3$ biological replicates). Data shown as mean \pm s.e.m. **d**, Co-localization of *Legionella*-containing vacuoles and RTN4 networks in primary mouse macrophages. Experiments in **a**, **b** and **d** were repeated twice independently with similar results. For gel source data, see Supplementary Fig. 1.



Extended Data Fig. 10 | Size-exclusion chromatography profile of SdeA_{FL}. SdeA_{FL} shows dimeric behaviour in size-exclusion chromatography column (Superdex 200 16/60). This experiment was

repeated twice independently with similar results. For the inset, $n = 1$. MW, molecular weight. For gel source data, see Supplementary Fig. 1.

Activity-dependent neuroprotective protein recruits HP1 and CHD4 to control lineage-specifying genes

Veronika Ostapczuk^{1,2}, Fabio Mohn¹, Sarah H. Carl^{1,3,5}, Anja Basters^{1,5}, Daniel Hess¹, Vytautas Iesmantavicius¹, Lisa Lampersberger^{1,4}, Matyas Flemlr¹, Aparna Pandey^{1,2}, Nicolas H. Thomä¹, Joerg Betschinger¹ & Marc Bühler^{1,2*}

De novo mutations in *ADNP*, which encodes activity-dependent neuroprotective protein (ADNP), have recently been found to underlie Helsmoortel–Van der Aa syndrome, a complex neurological developmental disorder that also affects several other organ functions¹. ADNP is a putative transcription factor that is essential for embryonic development². However, its precise roles in transcriptional regulation and development are not understood. Here we show that ADNP interacts with the chromatin remodeller CHD4 and the chromatin architectural protein HP1 to form a stable complex, which we refer to as ChAHP. Besides mediating complex assembly, ADNP recognizes DNA motifs that specify binding of ChAHP to euchromatin. Genetic ablation of ChAHP components in mouse embryonic stem cells results in spontaneous differentiation concomitant with premature activation of lineage-specific genes and in a failure to differentiate towards the neuronal lineage. Molecularly, ChAHP-mediated repression is fundamentally different from canonical HP1-mediated silencing: HP1 proteins, in conjunction with histone H3 lysine 9 trimethylation (H3K9me3), are thought to assemble broad heterochromatin domains that are refractory to transcription. ChAHP-mediated repression, however, acts in a locally restricted manner by establishing inaccessible chromatin around its DNA-binding sites and does not depend on H3K9me3-modified nucleosomes. Together, our results reveal that ADNP, via the recruitment of HP1 and CHD4, regulates the expression of genes that are crucial for maintaining distinct cellular states and assures accurate cell fate decisions upon external cues. Such a general role of ChAHP in governing cell fate plasticity may explain why ADNP mutations affect several organs and body functions and contribute to cancer progression^{1,3,4}. Notably, we found that the integrity of the ChAHP complex is disrupted by nonsense mutations identified in patients with Helsmoortel–Van der Aa syndrome, and this could be rescued by aminoglycosides that suppress translation termination⁵. Therefore, patients might benefit from therapeutic agents that are being developed to promote ribosomal read-through of premature stop codons^{6,7}.

ADNP contains nine N-terminal zinc-fingers and a C-terminal homeobox domain, strongly suggesting transcription factor activity⁸. Although originally associated with neuronal function⁹, ADNP is essential for embryonic development in mice: *Adnp*-deficient mouse embryos exhibit neural tube closure defects and die at days 8.5–9.5 of gestation. Two studies in knockout mouse embryos identified potential ADNP target genes that are implicated in cell differentiation and the maintenance of stem cells^{2,10}.

To dissect the molecular activity of ADNP, we exploited mouse embryonic stem (ES) cells¹¹. We first inserted a Flag-AviTag at the endogenous *Adnp* gene¹² (Extended Data Fig. 1a–c) and performed chromatin immunoprecipitation coupled to next-generation sequencing (ChIP-seq) to interrogate putative ADNP–DNA interactions genome-wide. This revealed 15,026 sites that are significantly enriched for ADNP (Fig. 1a, b and Supplementary Table 1). Notably, most (61%)

of the peaks were found in introns or proximal of annotated transcription start sites. The remaining peaks were located promoter distal in intergenic regions (Extended Data Fig. 1d, e). To analyse the function of ADNP, we generated homozygous *Adnp* knockout mouse ES cells (Extended Data Fig. 2a, b). Compared with wild-type ES cells, *Adnp*^{−/−} cells displayed gross morphological changes and appeared to differentiate spontaneously as they started spreading out of characteristically densely packed ES cell colonies (Fig. 1c, d). In addition, *Adnp*^{−/−} cells displayed heterogeneous activity of the pluripotency associated marker alkaline phosphatase (Fig. 1d). Transcriptome profiling by RNA sequencing (RNA-seq) revealed that most of the genes with altered mRNA levels in *Adnp*^{−/−} cells were upregulated (Extended Data Fig. 2c). Many genes bound by ADNP and displaying increased expression in the absence of ADNP encode known lineage specification factors, such as GATA4, GATA6, BMP1 or SOX17 (Supplementary Table 2). For example, *Gata4* is expressed predominantly in mesoderm- and endoderm-derived tissues¹³, and forced *Gata4* expression in mouse ES cells induces differentiation towards extra-embryonic endoderm¹⁴. Moreover, genes upregulated in *Adnp*^{−/−} cells were enriched for Gene Ontology terms related to differentiation and development (Extended Data Fig. 2d and Supplementary Table 3). We also observed a group of genes that were upregulated both in *Adnp*^{−/−} cells as well as in extra-embryonic endoderm stem-cell lines, which can be differentiated from mouse ES cells¹⁵ (Extended Data Fig. 2e). To gain further insight into the biological role of ADNP, we differentiated wild-type and *Adnp*^{−/−} ES cells towards neuronal precursor cells (Fig. 1c, d) using an established differentiation protocol¹⁶. *Adnp*^{−/−} ES cells formed smaller embryoid bodies and showed increased cell death after differentiation when compared to wild-type cells (Fig. 1d). *Nanog* and *Oct4* (also known as *Pou5f1*) expression was downregulated in both wild-type and *Adnp*^{−/−} cells, indicating successful exit from pluripotency (Fig. 1e). However, whereas *Adnp*^{+/+} cells started expressing neural markers such as *Pax6* and *Ng2* (also known as *Neurog2*) over the course of differentiation, *Adnp*^{−/−} cells failed to induce neural genes (Fig. 1f). Instead, the expression of *Gata4* and *Sox17* was specifically induced in *Adnp*^{−/−} cells (Fig. 1g), indicating misspecification towards the endodermal lineage under conditions that normally induce neuronal fate. This ES cell phenotype is reminiscent of *Adnp*^{−/−} mouse embryos, which show a developmental delay, fail to induce *Pax6* and suffer from defective neural tube closure². Thus, ADNP is required to restrain the expression of lineage-specifying genes in ES cells and for specification towards the neuronal lineage upon external differentiation cues.

These results are consistent with previously reported repressive activity of ADNP when artificially targeted to a reporter gene¹⁷. Furthermore, ADNP was shown to co-immunoprecipitate with the SWI/SNF chromatin remodelling complex¹⁸ or with proteins of the heterochromatin protein 1 (HP1) family^{10,19}. To unambiguously identify ADNP-interacting proteins in mouse ES cells, we subjected ADNP tagged endogenously with a Flag-AviTag to tandem-affinity purification coupled to liquid chromatography tandem mass spectrometry

¹Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland. ²University of Basel, Basel, Switzerland. ³Swiss Institute of Bioinformatics, Basel, Switzerland. ⁴University of Vienna, Vienna, Austria. ⁵These authors contributed equally: Sarah H. Carl, Anja Basters. *e-mail: marc.buehler@fmi.ch

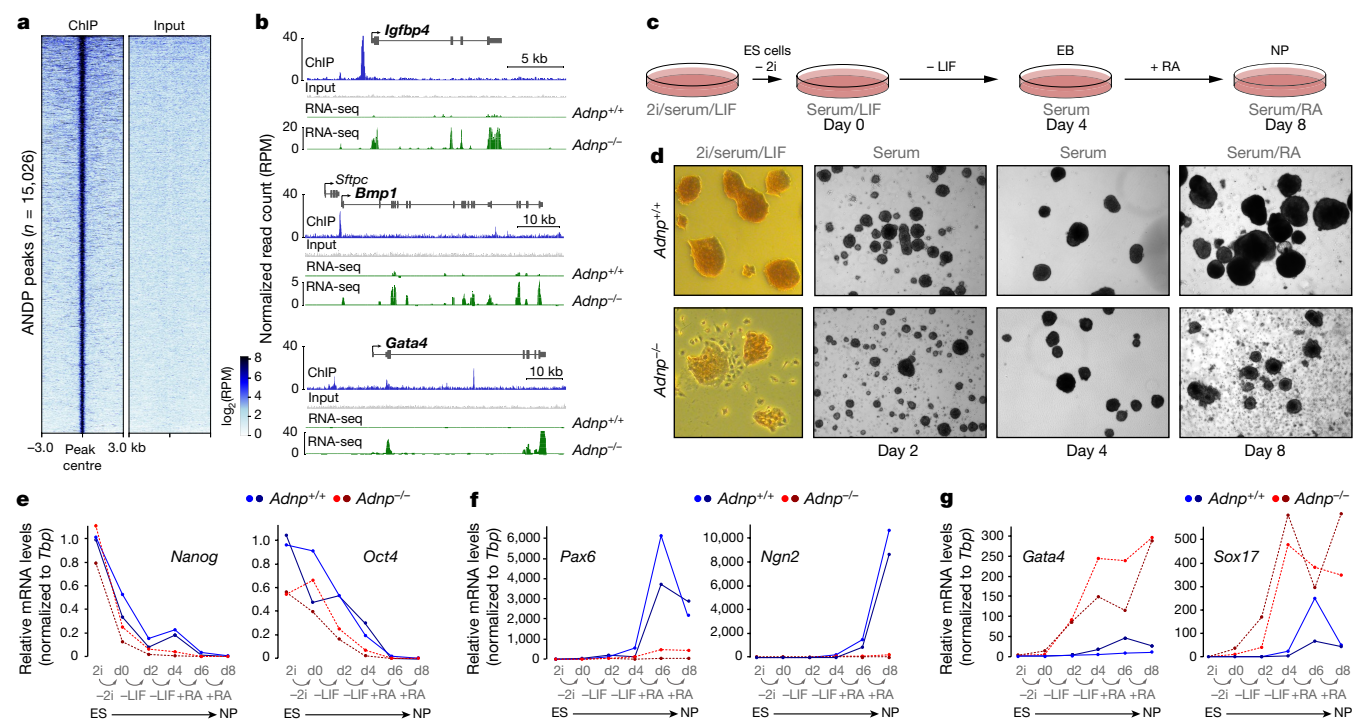


Fig. 1 | ADNP binds and represses lineage-specifying genes. **a**, Heat map of ADNP ChIP-seq enrichment across all significant peaks ($n = 15,026$) in the mouse genome. Each row represents a 6-kb window centred on peak midpoints, sorted by the ADNP ChIP signal. Input signals for the same windows are shown on the right. Average peak intensity of $n = 3$ biological replicates. RPM, reads per million. **b**, UCSC genome browser shots of three endoderm specification factors (*Igfbp4*, *Bmp1* and *Gata4*). ChIP-seq profiles for ADNP and input, and RNA-seq profiles for wild-type (*Adnp*^{+/+}) and ADNP-knockout (*Adnp*^{-/-}) mouse ES cells. Both ChIP-seq and RNA-seq profiles are normalized for library size. The experiment was repeated three times. **c**, Wild-type ES cells differentiate to neural progenitors (NP) in response to external cues. Consecutive withdrawal of 2i and leukemia inhibitory factor (LIF) results in the formation of

cellular aggregates (embryoid bodies, EB), which further differentiate into neural progenitors by the addition of retinoic acid (RA) at day 4. **d**, Phase-contrast images (original magnification, $\times 5$) of *Adnp*^{+/+} and *Adnp*^{-/-} mouse ES cells stained with alkaline phosphatase when grown in 2i, serum and LIF (day 0), and during differentiation towards the neuronal lineage (days 2, 4 and 8). The experiment was repeated three times. **e–g**, mRNA expression profiles of genes specifying pluripotent cells (**e**; *Nanog* and *Oct4*), or cells of the neural (**f**; *Pax6* and *Ngn2*) or the endodermal (**g**; *Gata4* and *Sox17*) lineages from two independent differentiation experiments (light and dark coloured dots). Values normalized to *Tbp* mRNA are shown relative to the *Adnp*^{+/+} parental cell line (in 2i/serum/LIF medium) for each replicate. Biological replicates were performed using independent mouse ES cell lines for each tagged protein. **d**, day.

(TAP-LC-MS/MS). Besides ADNP, we observed highly notable enrichment of HP1 β , HP1 γ and CHD4, but not SWI/SNF complex subunits. These interactions were preserved under 500 mM NaCl, showing that ADNP stably interacts with CHD4 and the HP1 β and HP1 γ proteins in ES cells (Fig. 2a). To corroborate this, we inserted a Flag-AviTag into the endogenous *Cbx1*, *Cbx3* and *Cbx5* genes, which encode the three mammalian HP1 isoforms HP1 β , HP1 γ and HP1 α , respectively¹² (Extended Data Fig. 3a). Both ADNP and CHD4 were highly enriched in HP1 β and HP1 γ purifications (Fig. 2b and Extended Data Fig. 3). By contrast, CHD4 did not co-purify with HP1 α (Extended Data Fig. 3b, e), and ADNP was 100-fold and 235-fold less abundant than in HP1 β and HP1 γ purifications, respectively (Extended Data Fig. 3g).

To verify that ADNP, HP1 and CHD4 form a stable complex via direct protein–protein interactions, we set out to reconstitute complex formation in vitro with recombinant human ADNP, HP1 γ and CHD4 (Extended Data Fig. 4). Co-lysis of cells expressing HP1 γ , ADNP and CHD4 resulted in the formation of a trimeric complex, which was preserved after streptavidin affinity purification, anion-exchange chromatography, and size-exclusion chromatography (SEC) (Fig. 2c, d). Subsequent experiments with full-length and truncated variants of the proteins (Extended Data Fig. 4) revealed that ADNP is at the core of the complex and interacts with CHD4 via its N terminus and with the chromoshadow domain (CSD) of HP1 via its C-terminal domain (Fig. 2e), probably through the PXXVL (in which X denotes any amino acid) motif¹⁷. In conclusion, CHD4, ADNP and HP1 β / γ form a stable protein complex, which we refer to as ChAHP.

Next we performed ChIP-seq with endogenously tagged HP1 α , HP1 β and HP1 γ and consulted a published dataset²⁰ for CHD4.

Corroborating the biochemistry, all of the ADNP peaks ($n = 15,026$, Fig. 1) showed enrichment for CHD4 and HP1 β / γ (Fig. 2f). Of the HP1 isoforms, the average HP1 γ occupancy was the highest, HP1 β was moderately enriched, and HP1 α was barely detectable at those sites (Fig. 2f and Extended Data Fig. 5a–d). This confirms our TAP-LC-MS/MS results and indicates that HP1 γ is the dominant isoform in ChAHP, whereas HP1 β is present in a minor fraction of ChAHP complexes or forms sub-stoichiometric heterodimers with HP1 γ . In line with a partial redundancy of HP1 β and HP1 γ , we observed that the average HP1 β occupancy on all ChAHP-bound sites was greatly increased in the absence of HP1 γ , whereas HP1 γ occupancy remained similar in the absence of HP1 β (Extended Data Fig. 5c). Thus, HP1 γ is the predominant member of ChAHP in ES cells.

HP1 proteins recognize and bind to methylated H3K9 through the chromodomain^{21,22}, indicating that HP1 might target ChAHP to H3K9 methylated nucleosomes. Consistent with previous immunostaining experiments¹⁷, we observed slight ADNP and CHD4 association with H3K9me3-marked chromatin. However, most of the highly enriched ADNP, and respective ChAHP peaks, were located in euchromatin (Fig. 2f). Consistent with repressive activity of ChAHP, histone modifications associated with active transcription were also absent (Extended Data Fig. 5e, f). In line with an H3K9me3-independent recruitment of ChAHP, HP1 γ with mutations in the chromodomain that abolish H3K9me binding still bound to ChAHP target genes (Extended Data Fig. 5g). By contrast, the binding of HP1 γ was lost at all ChAHP-bound sites in the absence of ADNP, whereas HP1 γ bound to genomic regions with H3K9me3-modified nucleosomes remained largely unaffected in *Adnp*^{-/-} cells (Fig. 3a). These results suggest that ADNP targets

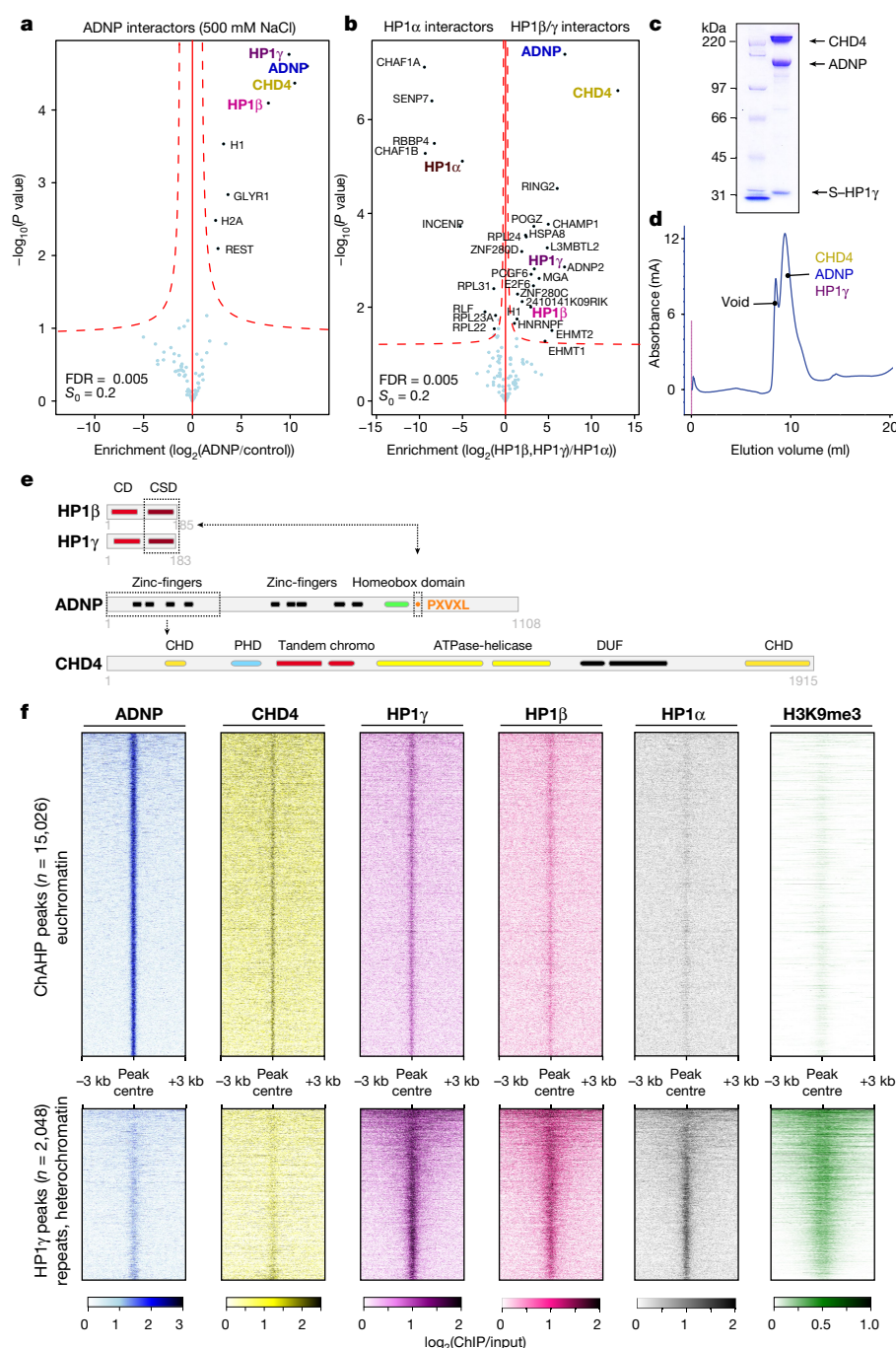


Fig. 2 | ADNP mediates ChAHP complex formation. **a**, TAP-MS/MS of ADNP endogenously tagged with Flag-AviTag. Protein purification was performed in the presence of 500 mM NaCl. Parental mouse ES cell line serves as background control. $n = 3$ biological replicates. FDR, false discovery rate. **b**, TAP-MS/MS of HP1α, HP1β and HP1γ endogenously tagged with Flag-AviTag. Protein purification was performed in the presence of 350 mM NaCl. Proteins that interact predominantly with HP1α (left), or HP1β or HP1γ (right) are indicated by UniProt names. $n = 3$ biological replicates. Statistical analysis was done with Perseus (see Methods). Mass spectrometry raw data are deposited with ProteomeXchange. **c**, **d**, In vitro reconstitution of the ChAHP complex. ADNP, CHD4 and HP1γ were expressed in Hi5 insect cells. Strep-tagged HP1γ (S-HP1γ) was pulled down with co-purifying ADNP and CHD4, followed by separation on size-exclusion chromatography (SEC) (Extended Data Fig. 4b, c). The fraction containing purified ChAHP was loaded on SDS-PAGE (**c**) and reinjected on SEC (**d**). For gel source data, see Supplementary Fig. 1. All experiments were performed at least twice. **e**, Scheme depicting ChAHP subunit interactions (see Extended Data Fig. 4). ADNP N-terminal zinc-fingers are necessary for the interaction with yet-to-be-determined CHD4 residues. The PXVXL motif in ADNP mediates the interaction with the CSD of HP1. Protein domains as predicted by InterPro. CD, chromodomain. **f**, Heat map of ADNP, CHD4, HP1γ, HP1β, HP1α and H3K9me3 ChIP-seq enrichment across all euchromatic sites bound by ChAHP (top) or heterochromatic sites bound by HP1γ (bottom). Each row represents a 6-kb window centred on the ADNP or HP1γ peak midpoint, respectively. Rows are sorted by ADNP (top) or HP1γ (bottom) ChIP enrichment. Average peak intensity of $n = 3$ biological replicates. Biological replicates were performed using independent mouse ES cell lines for each tagged protein.

the complex to euchromatic sites in a sequence-specific manner. Motif analysis of ADNP-bound loci revealed several significant DNA motifs. The highest-enriched motif (CGCCCYCTNSTG) was present in 63% of peaks ($P = 1 \times 10^{-10,538}$), and several motifs often co-occurred at bound genomic loci (Extended Data Fig. 6). To prove that ChAHP is indeed recruited via sequence-specific binding of ADNP, we deleted the predicted ADNP-binding motif (Δ motif) at the endogenous *Igfbp4* locus (Fig. 3b). Validating GCCCCTGGAG as an ADNP-binding site, ADNP enrichment was specifically lost at the *Igfbp4* locus in *Igfbp4* Δ motif/ Δ motif cells, whereas another ChAHP target gene (*Bmp1*) remained unaffected (Fig. 3c). Importantly, the binding of CHD4 and HP1γ was also depleted at the *Igfbp4* but not the *Bmp1* locus in *Igfbp4* Δ motif/ Δ motif cells (Fig. 3d, e). Consistent with ChAHP-mediated target repression, we observed significantly increased *Igfbp4* but not *Bmp1* mRNA levels in *Igfbp4* Δ motif/ Δ motif cells (Fig. 3f).

Identification of ChAHP strongly suggests that ADNP exerts its repressive function with the help of HP1. Indeed,

Cbx1 $^{-/-}$ *Cbx3* $^{-/-}$ *Cbx5* $^{-/-}$ triple-knockout cells revealed a distinct group of genes that was also upregulated in *Adnp* $^{-/-}$ cells. This was not evident in *Cbx* single- or double-knockout cells (Extended Data Fig. 7 and Supplementary Table 4). This suggests functional replacement by HP1α in the absence of HP1β and HP1γ, even though it only weakly interacts with ADNP and is not highly enriched at ChAHP target genes (Extended Data Figs. 3g, 5d). The fact that overall gene expression was not greatly affected if at least one HP1 isoform was present provides a general indication that HP1 isoforms can act partially redundantly to repress target genes in ES cells (Extended Data Fig. 7b).

The requirement of HP1 for ChAHP-mediated repression prompted us to revisit ADNP mutations found in patients with Helsmoortel-Van der Aa syndrome. Most are frameshift or nonsense mutations that result in C-terminally truncated ADNP that lacks the homeobox domain and the HP1 interaction motif¹ (Extended Data Fig. 1b). This suggests that mutant ADNP fails to assemble functional ChAHP and/or to bind its target genes. To test this, we introduced a patient-specific nonsense

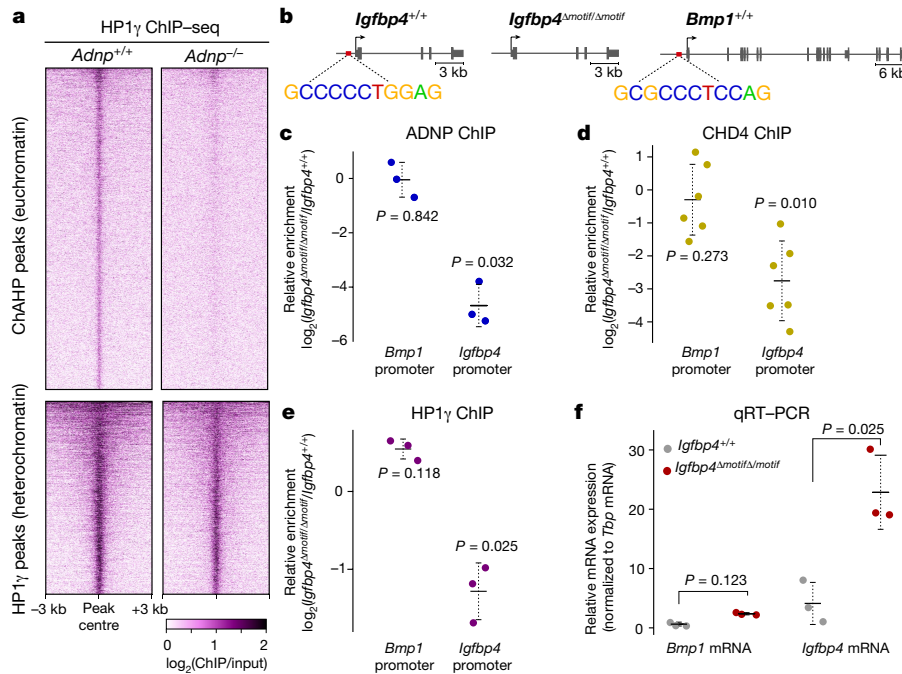


Fig. 3 | DNA sequence specifies ChAHP association with euchromatin.

a, Heat map of HP1 γ ChIP-seq enrichment in euchromatic ChAHP-bound sites (top) or heterochromatic HP1-bound sites (bottom) in *Adnp*^{+/+} and *Adnp*^{-/-} mouse ES cells. Each row represents a 6-kb window centred on the respective peak signal. Rows are sorted by mean ChIP enrichment. Average peak intensity of *n* = 3 biological replicates (that is, three independent ES cell lines). **b**, Schemes depicting the location of ADNP-binding motifs in the *Igfbp4* (left) and *Bmp1* (right) genes and the *Igfbp4* locus with the motif deletion *Igfbp4* ^{Δ motif/ Δ motif} (middle). **c**, ChIP-qPCR measuring ADNP enrichments at *Igfbp4* and *Bmp1* promoters in

Igfbp4 ^{Δ motif/ Δ motif} cells compared to the parental line (*Igfbp4*^{+/+}). *n* = 3 biological replicates. **d**, ChIP-qPCR measuring CHD4 enrichments at *Igfbp4* and *Bmp1* promoters in *Igfbp4* ^{Δ motif/ Δ motif} cells compared to the parental line. *n* = 6 biological replicates. **e**, ChIP-qPCR measuring HP1 γ enrichments at *Igfbp4* and *Bmp1* promoters in *Igfbp4* ^{Δ motif/ Δ motif} cells compared to the parental line. *n* = 3 biological replicates. **f**, qRT-PCR measurement of *Igfbp4* and *Bmp1* mRNA levels in wild-type and *Igfbp4* ^{Δ motif/ Δ motif} cells. *n* = 3 biological replicates. *P* values in **c**–**f** were calculated using two-tailed unpaired unequal variances *t*-tests. Centre values denote the mean; error bars denote s.d.

mutation upstream of the homeobox domain at amino acid position 718 (*Adnp*^{PTC718}; Extended Data Fig. 8a; corresponds to Tyr719 in human ADNP). ADNP^{PTC718} failed to co-purify with HP1 β and HP1 γ , whereas the interaction with CHD4 remained preserved (Extended Data Fig. 8b). Consistent with the requirement of HP1 for silencing, we observed increased expression of ADNP-target genes in cells that express ADNP^{PTC718} (Extended Data Fig. 8c). The ADNP^{PTC718} protein still bound its target site at the *Igfbp4* locus (Extended Data Fig. 8d), indicating that the homeobox domain is dispensable for DNA binding but might assist in target repression. These results demonstrate that patients with nonsense mutations in the *ADNP* gene cannot assemble fully functional ChAHP complexes. To test whether this could potentially be restored pharmacologically, we treated ADNP^{PTC718}-expressing cells with gentamycin or paromomycin, two aminoglycoside antibiotics that promote translational read-through⁵. Indeed, gentamycin treatment promoted read-through of PTC718 (Extended Data Fig. 8e, f) and rescued the interaction between ADNP^{PTC718} and HP1 β or HP1 γ . Although less effectively, HP1 γ was also retrieved from samples that were treated with paromomycin (Extended Data Fig. 8b). Thus, a therapeutic approach that promotes ribosomal read-through of premature stop codons could be of considerable medical benefit. However, the discovery of new nonsense suppressors will be inevitable, because the clinical utility of aminoglycoside therapy is limited by low efficacy and serious toxicities⁶.

Finally, we set out to investigate the molecular function of ChAHP. Inspired by the role of CHD4 in nucleosome remodelling and HP1 in heterochromatin assembly, we investigated a possible role of ChAHP in regulating local chromatin accessibility by the ATAC-seq method (assay for transposase-accessible chromatin using sequencing)²³. Many transcription factors, such as NRF1²⁴, generate local accessible regions at their DNA-binding sites (Fig. 4a). Unexpectedly, we did not observe such footprints for ADNP. Instead, chromatin across ChAHP-bound

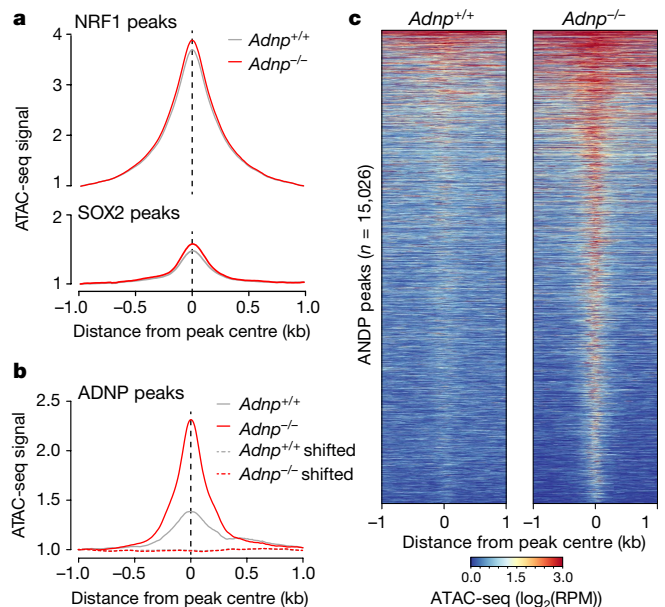


Fig. 4 | ChAHP obstructs chromatin accessibility. **a**, Average accessibility of loci bound by unrelated transcription factors NRF1 (top) and SOX2 (bottom) in *Adnp*^{+/+} (grey) and *Adnp*^{-/-} (red) mouse ES cell lines measured by ATAC-seq. Profiles represent averaged biological replicates (*n* = 4). **b**, Average accessibility of loci bound by ADNP and ‘random’ control loci (ADNP peaks shifted by 10 kb; dashed lines) in *Adnp*^{+/+} (grey) and *Adnp*^{-/-} (red) mouse ES cells. Profiles represent averaged biological replicates (*n* = 4). **c**, Heat map showing ATAC-seq read coverage in a 2-kb window around all ADNP peaks normalized by library depth in *Adnp*^{+/+} (left) and *Adnp*^{-/-} (right) ES cells.

loci was largely devoid of an ATAC-seq signal (Fig. 4b, c). This suggests that ChAHP is bound to chromatin with impaired accessibility, or conversely, that the binding of ChAHP renders chromatin inaccessible. Notably, all ChAHP-bound sites became readily accessible in the absence of ADNP, whereas ChAHP-independent control loci such as NRF1- or SOX2-binding sites, as well as a 'random' set of genomic loci (ADNP peaks shifted by 10 kb) showed no difference in accessibility (Fig. 4a–c). Notably, the opening of chromatin in *Adnp*^{−/−} cells was restricted to a few hundred base pairs around ChAHP-binding sites, and the surrounding regions remained inaccessible (Fig. 4c). Thus, rather than assembling broad, inaccessible domains of chromatin, ChAHP denies direct access to its cognate DNA-binding sites.

In summary, we have discovered ChAHP, a gene-regulatory complex that consists of the chromatin remodeller CHD4, the DNA-binding factor ADNP, and the heterochromatin proteins HP1 β and HP1 γ . By locally restricting access to DNA, ChAHP prevents endodermal gene transcription in mouse ES cells and during neuroectodermal differentiation. This stabilizes cellular states and ensures correct lineage specification. Although ChAHP could directly interfere with transcribing RNA polymerase, we favour a model in which ChAHP prevents the binding of other regulatory factors, such as transcriptional activators, to DNA. Although the exact mode of action of ChAHP remains to be determined, such a model would be consistent with the notion that ChAHP also binds outside gene bodies and promoters.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0153-8>.

Received: 6 September 2017; Accepted: 13 April 2018;

Published online 23 May 2018.

- Helsmoortel, C. et al. A SWI/SNF-related autism syndrome caused by de novo mutations in ADNP. *Nat. Genet.* **46**, 380–384 (2014).
- Pinhasov, A. et al. Activity-dependent neuroprotective protein: a novel gene essential for brain formation. *Brain Res. Dev. Brain Res.* **144**, 83–90 (2003).
- Gozes, I., Yeheskel, A. & Pasmanik-Chor, M. Activity-dependent neuroprotective protein (ADNP): a case study for highly conserved chordata-specific genes shaping the brain and mutated in cancer. *J. Alzheimers Dis.* **45**, 57–73 (2015).
- Gozes, I. et al. The compassionate side of neuroscience: Tony Sermone's undiagnosed genetic journey—ADNP mutation. *J. Mol. Neurosci.* **56**, 751–757 (2015).
- Hermann, T. Aminoglycoside antibiotics: old drugs and new therapeutic approaches. *Cell. Mol. Life Sci.* **64**, 1841–1852 (2007).
- Peltz, S. W., Morsy, M., Welch, E. M. & Jacobson, A. Ataluren as an agent for therapeutic nonsense suppression. *Annu. Rev. Med.* **64**, 407–425 (2013).
- Welch, E. M. et al. PTC124 targets genetic disorders caused by nonsense mutations. *Nature* **447**, 87–91 (2007).
- Zamostiano, R. et al. Cloning and characterization of the human activity-dependent neuroprotective protein. *J. Biol. Chem.* **276**, 708–714 (2001).
- Bassan, M. et al. Complete sequence of a novel protein containing a femtomolar-activity-dependent neuroprotective peptide. *J. Neurochem.* **72**, 1283–1293 (1999).
- Mandel, S., Rechavi, G. & Gozes, I. Activity-dependent neuroprotective protein (ADNP) differentially interacts with chromatin to regulate genes essential for embryogenesis. *Dev. Biol.* **303**, 814–824 (2007).
- Niwa, H. Mouse ES cell culture system as a model of development. *Dev. Growth Differ.* **52**, 275–283 (2010).
- Flemr, M. & Bühler, M. Single-step generation of conditional knockout mouse embryonic stem cells. *Cell Reports* **12**, 709–716 (2015).
- Molkentin, J. D. The zinc finger-containing transcription factors GATA-4, -5, and -6. Ubiquitously expressed regulators of tissue-specific gene expression. *J. Biol. Chem.* **275**, 38949–38952 (2000).
- Fujikura, J. et al. Differentiation of embryonic stem cells is induced by GATA factors. *Genes Dev.* **16**, 784–789 (2002).
- Cho, L. T. Y. et al. Conversion from mouse embryonic to extra-embryonic endoderm stem cells reveals distinct differentiation capacities of pluripotent stem cell states. *Development* **139**, 2866–2877 (2012).
- Bibel, M. et al. Differentiation of mouse embryonic stem cells into a defined neuronal lineage. *Nat. Neurosci.* **7**, 1003–1009 (2004).
- Mosch, K., Franz, H., Soeroes, S., Singh, P. B. & Fischle, W. HP1 recruits activity-dependent neuroprotective protein to H3K9me3 marked pericentromeric heterochromatin for silencing of major satellite repeats. *PLoS ONE* **6**, e15894 (2011).
- Mandel, S. & Gozes, I. Activity-dependent neuroprotective protein constitutes a novel element in the SWI/SNF chromatin remodeling complex. *J. Biol. Chem.* **282**, 34448–34456 (2007).
- Vermeulen, M. & et al. Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. *Cell* **142**, 967–980 (2010).
- de Dieuleveult, M. et al. Genome-wide nucleosome specificity and function of chromatin remodellers in ES cells. *Nature* **530**, 113–116 (2016).
- Bannister, A. J. et al. Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410**, 120–124 (2001).
- Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410**, 116–120 (2001).
- Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
- Domcke, S. et al. Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* **528**, 575–579 (2015).

Acknowledgements We thank A. Peters and D. Schübeler for discussion and feedback on the manuscript, and H. Pickersgill for editing assistance. We thank Y. Shimada and N. Laschet for technical support. We would like to thank the FMI Functional Genomics facility for assistance in library construction and next generation sequencing, and J. Seebacher for discussions. This work was supported by funds from the Swiss National Science Foundation (SNF).

Reviewer information Nature thanks P. Wade and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions V.O. designed and performed experiments, analysed data, generated cell lines and prepared figures. M.F. generated cell lines and advised on experimental design. D.H. and V.I. acquired and analysed mass spectrometry data. Bioinformatic and computational analysis was performed by S.H.C. and F.M. A.B. performed in vitro biochemistry experiments. V.O. and L.L. performed immunoprecipitation and mass spectrometry with patient-specific *Adnp* mutants, and L.L. performed ChIP experiments for the patient-specific *Adnp* mutants. A.P. performed ChIP experiments for the *Cbx3* chromodomain mutant. F.M. generated and analysed ATAC-seq data. N.H.T. and J.B. advised on the experimental design. M.B. conceived and supervised the study, and secured funding. M.B. and F.M. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Competing interests The authors declare competing financial interests: a patent application has been filed (EP17191642.2). The Friedrich Miescher Institute for Biomedical Research (FMI) receives significant financial contributions from the Novartis Research Foundation. Published research reagents from the FMI are shared with the academic community under a Material Transfer Agreement (MTA) having terms and conditions corresponding to those of the UBMTA (Uniform Biological Material Transfer Agreement).

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0153-8>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0153-8>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Cell culture and genome editing. Mouse ES cells (129×C57BL/6²⁵) were cultured on gelatin-coated dishes in ES medium containing DMEM (Gibco 21969-035), supplemented with 15% fetal bovine serum (FBS; Gibco), 1× non-essential amino acids (Gibco), 1 mM sodium pyruvate (Gibco), 2 mM L-glutamine (Gibco), 0.1 mM 2-mercaptoethanol (Sigma), 50 mg ml⁻¹ penicillin, 80 mg ml⁻¹ streptomycin, 3 μM glycogen synthase kinase (GSK) inhibitor (Calbiochem, D00163483), 10 μM MEK inhibitor (Tocris, PD0325901), and homemade LIF, at 37 °C in 5% CO₂.

The genome editing was performed as previously published¹² in the absence of GSK and MEK inhibitors in the above-described ES medium.

Generation of endogenously tagged ES cell lines. For endogenous gene tagging using TALENs, Rosa26:Flag-V5-expressing cells (cMB053 or cMB063) were transfected with 400 ng TALEN-EED, 400 ng TALEN-KKR, 100 ng pRRP reporter and 1,000 ng of donor single-stranded oligodeoxynucleotide encoding the tag sequence. The single-stranded oligodeoxynucleotides were synthesized as Ultramers by Integrated DNA Technologies and their sequences are listed along with TALEN sequences in Supplementary Table 6. All transfections were carried out using Lipofectamine 3000 reagent (Invitrogen) at a 3 μl:1 μg DNA ratio in OptiMEM medium (Invitrogen). Transfected cells were selected by adding puromycin (2 μg ml⁻¹) to the ES medium 24 h after transfection. After 36 h of selection, surviving cells were sparsely seeded for clonal expansion. The resulting clones were individually picked, split and screened by western blot for desired tag integration. See Supplementary Information for the list of tagged cell lines generated in this study.

Straight knockout ES cell line generation. *Cbx1*^{-/-} mouse ES cells were generated using TALENs that target the first and last coding exon, resulting in a deletion of approximately 6,000 bp (exon 2–exon 6).

Adnp^{-/-} mouse ES cells were generated using *Cas9* and TALENs that target the first and last coding exon, resulting in a deletion of approximately 7,000 bp (exon 2–exon 4). The *Cas9* sgRNA sequence was cloned into the SpCas9-2A-mCherry plasmid²⁶. Sequences of TALENs and *Cas9* sgRNA can be found in Supplementary Table 6. See Supplementary Information for the list of straight knockout cell lines generated in this study.

Conditional ES cell line generation. The *Cbx3*^{fl/fl} cell line was generated as described¹². For the *Cbx5*^{fl/fl} conditional cell line, a mouse ES cell line containing an integration of the CreERT2 recombinase fusion in the *Rosa26* locus (cMB052 or cMB063) was transfected with TALENs cutting before and after the third exon. Single-stranded oligodeoxynucleotides with corresponding homology arms and *loxP* sites for integration were also included in the transfection mix (see Supplementary Table 6 for sequences). Clones were screened for homozygous integrations for both *loxP* sites. A cell line with both bi-allelic *loxP* integrations was tested for recombination efficiency by treating the cells with 0.1 μM 4-hydroxytamoxifen (4-OHT; Sigma) followed by western blot or quantitative PCR with reverse transcription (qRT-PCR).

Transient expression experiments in ES cells. The full-length *Adnp* cDNA was cloned into the mammalian expression vector pEFaFB (promoter of elongation factor-1 alpha; ATG-3×Flag-Avi-GOI-2A-puromycin), creating pEFaFB-*Adnp*, which was then used as a template to mutate codon 718 (TAT to TAA) using the QuikChange Lightning Site-Directed Mutagenesis (SDM) Kit (Agilent), with the final construct encoding pEFaFB-ADNP^{PTC718}. Alternatively, *Adnp*^{PTC718} cDNA was cloned into the pEFaCFB vector (C-terminal 3×Flag-AviTag).

The *Cbx3* cDNA was cloned into pEFaFB, and then used as a template for SDM mutating chromodomain residues Trp43 and Phe46 to Ala (*Cbx3* CDMut)²⁷.

For ChIP experiments, 5 × 10⁶ *Adnp*^{-/-} (cMB377) cells in a 10-cm dish were reverse transfected with 10 μg pEFaFB-*Adnp* or pEFaFB-*Adnp*^{PTC718}. Alternatively, *Cbx1*^{-/-}*Cbx3*^{-/-}*Cbx5*^{-/-} triple-knockout (4-OHT-treated cMB282) cells were transfected with 10 μg pEFaFB-*Cbx3* or pEFaFB-*Cbx3*-CDmut. Cells were collected 48 h after transfection, and further processed according to the ChIP protocol (below). For affinity purification experiments, 6 × 10⁶ cells in a 15-cm dish were reverse transfected with 10 μg pEFaFB-*Adnp*^{PTC718} or pEFaCFB-*Adnp*^{PTC718}. Twenty hours after transfection, cells were forward-transfected with 5 μg pEFaFB-*Adnp*^{PTC718}. Finally, 48 h from the first transfection, cells were treated with 2 mg ml⁻¹ gentamycin (Sigma, G1914) or paromomycin (Sigma, P9297) for 24 h. Cells were then collected and processed according to the Affinity purification protocol (below). Transfections were carried out using Lipofectamine 3000 reagent (Invitrogen) at a 3 μl:1 μg DNA ratio in OptiMEM medium (Invitrogen), and 10 μM ROCK inhibitor (Tocris, Y27632) for increased cell survival.

Differentiation of ES cells to neuronal precursors. The cMB263 (*Adnp*^{+/+}) and cMB267 (*Adnp*^{-/-}) ES cell lines were differentiated as previously described²⁸, except that no feeder cells were used. Instead, cells were grown in ES cell medium containing 2i, as described above.

Western blotting. Cells were grown to confluency on 6-well plates, collected in PBS, pelleted by 2 min centrifugation at 400g, and pellets were then resuspended in 100 μl protein extraction buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1% Triton X-100, 0.5 mM EDTA, and 5% glycerol) supplemented with protease inhibitor cocktail (PIC; Roche), 1 mM PMSF, and 1 mM dithiothreitol (DTT). Proteins were extracted for 30 min on ice, the lysates were centrifuged at 16,000g for 20 min at 4 °C, and the protein concentration in the supernatant was determined using the BioRad protein assay. For western blotting, 20 μg of protein was resolved on NuPAGE-Novex Bis-Tris 4–12% gradient gels (Invitrogen), which were semi-dry transferred on polyvinylidene fluoride (PVDF) membrane, blocked for 30 min in 2.5% non-fat dry milk in TBS plus 0.05% Tween 20 (TBST), and stained with primary antibodies at 4 °C overnight. The primary antibodies used for western blotting were mouse anti-Flag (1:1,000, Sigma clone M2), goat-anti-HP1α (1:1,000, Abcam, ab77256), mouse-anti-HP1α (1:1,000, Millipore, mab3446), rat-anti-HP1β (1:500, Serotec, MCA1946), mouse-anti-HP1γ (1:2,000, Cell Signaling Technology), mouse-anti-CHD4 (1:1,000, Abcam, ab70469), rabbit-anti-MTA2 (1:1,000, Bethyl, A300-395A-T), rabbit-anti-GATAD2B (1:1,000, Bethyl, A301-283A-T), rabbit-anti-MBD3 (1:1,000, Bethyl, A302-528A-T) and rat-anti-tubulin (1:5,000, Abcam clone YL1/2). Signal was detected with corresponding horseradish peroxidase (HRP)-conjugated secondary antibodies and Immobilon Western Chemiluminescent HRP Substrate (Millipore). For streptavidin staining, membranes were blocked after transfer in 2% bovine serum albumin (BSA) in TBST and incubated with streptavidin–HRP (1:20,000, Sigma) for 30 min at room temperature, followed by signal development as above.

ChIP. A confluent 10-cm culture dish of ES cells (approximately 2 × 10⁷ cells) were cross-linked for 7 min at room temperature, with 1% formaldehyde solution (Sigma, F8775) added directly to the ES medium. Cross-linking was quenched by the addition of glycine to a final concentration of 0.125 M and incubation at 4 °C for 10 min; cells were then washed twice with PBS. Cells were collected in 1 ml PBS with PIC (Roche) and spun at 600g for 5 min at 4 °C. Cells were then resuspended in 5 ml wash solution I (10 mM Tris pH 8, 10 mM EDTA, 0.5 mM EGTA, 0.25% Triton X-100), incubated for 10 min on ice, then spun at 1,200g for 5 min at 4 °C. The remaining nuclear pellet was then resuspended in 5 ml wash solution II (10 mM Tris pH 8, 1 mM EDTA, 0.5 mM EGTA and 200 mM NaCl) and incubated for 5 min on ice, then spun at 1,200g for 5 min at 4 °C. Cell pellet was subsequently washed in 900 μl sonication buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA and 0.1% SDS) without disturbing the pellet, and finally resuspended in the sonication buffer supplemented with PIC. Chromatin was then sonicated in Covaris 1-ml tubes for 15 min with the following settings: duty cycle: 5%, peak incident power: 140 W, cycles per burst: 200, temperature (bath): 4 °C.

Beads preparation. For Bio-ChIP (ChIP for proteins tagged with the Flag-Avi tag), 40 μl Dynabeads Stepavidin (Thermo Fisher) per sample, or alternatively 40 μl Protein-G Dynabeads (Thermo Fisher) per sample for ChIP with protein-specific antibodies (Ab-ChIP), were washed twice for 5 min in 0.5 ml blocking buffer (PBS, 0.5% Tween and 0.5% BSA). Streptavidin Dynabeads were then washed twice with immunoprecipitation buffer (50 mM HEPES pH 7.5, 150 mM NaCl, 1 mM EDTA, 0.1% SDS, 0.1% sodium deoxycholate and 1% Triton X-100) and stored on ice. Protein-G Dynabeads were incubated for 1 h at room temperature in blocking buffer with the desired antibody. Beads were then washed twice in blocking buffer and stored on ice. For CHD4 ChIP, 10 μg mouse anti-CHD4 (Abcam, ab70469, 3F2/4) conjugated to Protein G was used.

Immunoprecipitation and washes. For immunoprecipitation analyses, 10 μl (1%) was kept as the input sample, and 40 μl pre-blocked Dynabeads were added to 1 ml of sonicated chromatin in immunoprecipitation buffer and incubated overnight at 4 °C on a rotating wheel. Beads were collected on a magnetic rack for 2–3 min to remove supernatant between each step, and washed as follows: for Bio-ChIP, twice for 10 min with 2% SDS in TE buffer (10 mM Tris, pH 8, 1 mM EDTA), once for 10 min with high salt buffer (50 mM HEPES pH 7.5, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate and 500 mM NaCl), once for 10 min with DOC buffer (250 mM LiCl, 0.5%, NP-40, 0.5% deoxycholate, 1 mM EDTA and 10 mM Tris pH 8) and twice for 10 min with 1 ml TE buffer. For Ab-ChIP, beads were washed five times with immunoprecipitation buffer, twice with DOC buffer, and twice with TE buffer. Beads were then resuspended in 300 μl elution buffer (1% SDS and 100 mM NaHCO₃) and 6 μl RNaseA (10 mg ml⁻¹ stock) and incubated at 37 °C for 30 min while mixing. Elution buffer was adjusted with 6 μl 0.5 M EDTA, 12 μl 1 M Tris pH 8 and 2.5 μl Proteinase K (10 mg ml⁻¹, Roche). Beads were incubated for 3 h at 55 °C and then overnight at 65 °C with mixing to de-crosslink. The same procedure was followed for input samples including RNase and proteinase K digestion. DNA was purified using AMPure XP beads (Beckman Coulter). Quantification was performed using Qubit dsDNA high-sensitivity assay (Thermo Fisher).

ChIP-qPCR and ChIP-seq. DNA was subjected to qPCR analysis (as described for qRT-PCR, below) using ChIP primers described in Supplementary Information. For ChIP-seq sample preparation, library construction was performed using the NEBNext Ultra kit (New England Biolabs) following

manufacturer recommendations. Libraries were sequenced on Illumina HiSeq 2500 machines, with 50-bp single-end sequencing.

qRT-PCR and RNA-seq. For qRT-PCR experiments, total RNA was extracted from ES cells with the Absolutely RNA Microprep Kit (Stratagene). Total RNA (500 ng) was reverse transcribed with the Primescript RT kit (Clontech). qRT-PCR was performed on a CFX96 Real-Time PCR System (Bio-Rad) using the SsoAdvanced SYBR Green Supermix (Bio-Rad, 172–5264). Relative RNA levels were calculated from C_t values according to the ΔC_t method and normalized to *Tbp* mRNA levels where applicable. For RNA-seq, total RNA (isolated as above) was subjected to ribosomal RNA depletion using the Ribozero kit (Illumina) followed by library construction using the ScriptSeq V2 library preparation kit (Illumina).

ATAC-seq. ATAC-seq was performed following a previously published protocol²³ using 50,000 *Adnp*^{+/+} or *Adnp*^{-/-} mouse ES cells. The experiment was performed in biological replicates using two independent isogenic cell lines for each genotype. Libraries were paired-end sequenced (2 × 75 bp) using an Illumina NextSeq 500 device.

Affinity purification for LC-MS/MS. All affinity purifications in this work were performed according to the following protocol with the exception of ADNP^{PTC718} affinity purifications (see below). Cells were grown to confluency on 10-cm dishes, collected in PBS, and pelleted by centrifugation at 400g for 2 min. All subsequent steps were performed on ice or at 4 °C. Pellets were resuspended in 3 ml of nuclear extract buffer 1 (NEB1; 20 mM HEPES, 10 mM KCl, 1 mM EDTA, 0.1 mM Na₃VO₄, 0.2% NP-40, 10% glycerol, 1 mM DTT and 1 × PIC) followed by centrifugation at 1,000g for 3 min. Pellets were resuspended in 1 ml NEB1 buffer and incubated on ice for 10 min, followed by dounce homogenization. Isolated nuclei were collected by centrifugation at 1,000g for 15 min, and carefully washed twice with 1 ml NEB1 without disturbing the pellet. Pellets were then resuspended in 0.5 ml of nuclear extract buffer 2 (NEB2; 20 mM HEPES, 10 mM KCl, 1 mM EDTA, 0.1 mM Na₃VO₄, 350 mM NaCl, 20% glycerol, 1 mM DTT and 1 × PIC), dounce homogenized (20 × up and down), incubated for 30 min, and finally spun at 16,000g for 30 min. Protein concentration was determined using the Bradford assay, and approximately 250 µg of nuclear extract was used per affinity purification. The protein lysates were adjusted to affinity purification buffer (350 mM or 500 mM NaCl, 20 mM Tris-HCl, pH 7.5, 0.3% NP-40, 1 mM EDTA, 10% glycerol, 1 mM DTT and 1 × PIC), added to 20 µl anti-Flag-M2 Dynabeads (Sigma), and incubated overnight rotating at 4 °C. Dynabeads were washed the next day in affinity purification buffer (4 × 10 min), followed by 3 × 15-min elutions of bound proteins with 3 × Flag peptide (final concentration 0.3 mg ml⁻¹ in affinity purification buffer, Sigma). Next, elutions were pooled and added to the washed Streptavidin Dynabeads (Thermo Fisher), and incubated overnight rotating at 4 °C. Streptavidin Dynabeads were washed the next day with affinity purification buffer (4 × 10 min), followed by a wash with affinity purification buffer without NP40. For single-step affinity purification, Flag purification was omitted, and lysates were directly applied to the Streptavidin Dynabeads. The enriched proteins were digested directly on the Dynabeads with 0.1 mg ml⁻¹ trypsin in digestion buffer (50 mM Tris pH 8.0, 1 mM CaCl₂ and 1 mM TCEP).

For ADNP^{PTC718} affinity purification, cells from 2 × 15-cm dishes per replicate were used; collection and nuclear lysate isolation were as described above. Next, 400 µg of nuclear lysates were used for single-step purification with Streptavidin Dynabeads in affinity purification buffer (2 h incubation at 4 °C), followed by washes (see above). The enriched proteins were digested directly on the Dynabeads with 0.2 µg Lys-C in 5 µl digestion buffer (3 M guanidium chloride, 20 mM EPPS, pH 8.5, 10 mM CAA and 5 mM TCEP) for 2 h at room temperature. Next, samples were diluted with 50 mM HEPES, pH 8.5, and digested with 0.2 µg trypsin overnight at 37 °C. The next day, 0.2 µg fresh trypsin was added, and samples were incubated for an additional 5 h at 37 °C.

Mass spectrometry. Analysis of affinity purification. The generated peptides (see 'Affinity purification for LC-MS/MS') were acidified with TFA to a final concentration of 0.8% and analysed by LC-MS/MS with an EASY-nLC 1000 using the two-column set-up (Thermo Scientific). The peptides were loaded with 0.1% formic acid, 2% acetonitrile in H₂O onto a peptide trap (Acclaim PepMap 100, 75 µm × 2 cm, C18, 3 µm, 100 Å) at a constant pressure of 80 MPa. Peptides were separated, at a flow rate of 150 nl min⁻¹ with a linear gradient of 2–6% buffer B in buffer A in 3 min followed by an linear increase from 6 to 22% in 40 min, 22–28% in 9 min, 28–36% in 8 min, 36–80% in 1 min and the column was finally washed for 14 min at 80% buffer B in buffer A (buffer A: 0.1% formic acid; buffer B: 0.1% formic acid in acetonitrile) on a 50 µm × 15 cm ES801 C18, 2 µm, 100 Å column (Thermo Scientific) mounted on a DPV ion source (New Objective) connected to a Orbitrap Fusion (Thermo Scientific). The data were acquired using 120,000 resolution for the peptide measurements in the Orbitrap and a top T (3 s) method with HCD fragmentation for each precursor and fragment measurement in the ion trap according to the recommendation of the manufacturer (Thermo Scientific).

Protein identification and relative quantification of the proteins was done with MaxQuant version 1.5.3.8 using Andromeda as search engine²⁹ and label-free quantification (LFQ³⁰) as described previously³¹. The mouse subset of the UniProt version 2015_01 combined with the contaminant DB from MaxQuant was searched and the protein and peptide FDR values were set to 0.01. All MaxQuant parameters can be found in the uploaded parameterfile: mqpar.xml (deposited in the PRIDE repository, see Data availability).

Statistical analysis was done in Perseus (version 1.5.2.6)^{29,30,32}. Results were filtered to remove reverse hits, contaminants and peptides found in only one sample. Missing values were imputed and potential interactors were determined using *t*-test and visualized by a volcano plot. Significance lines corresponding to a given FDR have been determined by a permutation-based method³³. Threshold values (FDR) were selected between 0.005 and 0.05 and *S*₀ (curve bend) between 0.2 and 2.0 and are shown in the corresponding figures. Results were exported from Perseus and visualized using statistical computing language R.

iBAQ. Intensity based absolute quantification (iBAQ) was done as described previously³⁴ to evaluate protein abundances in the ChAHP complexes of the different pull-down reactions.

PRM data acquisition. Parallel reaction monitoring (PRM) analyses were performed using the same LC-MS system and gradient as described above. The acquisition method consisted of acquiring one MS spectrum at 120,000 resolution from 375 to 1,575 Da followed by 21 PRM spectra. An isolation window of 1.6 Da, a resolution of 240,000, and an automatic gain control value of 10⁵ was used. Fragmentation was performed with a stepped HCD collision energy of 30 ± 5, and MS/MS scans were acquired with a scan range from 110 to 1,800.

PRM data analysis. The acquired PRM data were processed using Skyline 4.1³⁵. The transition selection was systematically verified and adjusted when necessary to ensure that no co-eluting contaminant distorted quantification based on traces co-elution (retention time) and the correlation between the relative intensities of the endogenous fragment ion traces, and their counterparts from the library.

MASCOT 2.5 was used in the Decoy mode to search the Swissprot mouse version 2015_01 including common contaminants. The enzyme specificity was set to trypsin allowing for up to three incomplete cleavage sites. Carbamidomethylation of cysteine (+57.0245) was set as a fixed modification, oxidation of methionine (+15.9949 Da) and acetylation of the protein N terminus (+42.0106 Da) were set as variable modifications. Parent ion mass tolerance was set to 10 p.p.m. and fragment ion mass tolerance to 0.6 Da. The results were validated with the program Scaffold Version 4.4 (Proteome Software). Protein identifications were accepted if they could be established at greater than 0.1% FDR rate as calculated in Scaffold.

SEC of nuclear lysates. Nuclear lysates were isolated as described above ('Affinity purification for LC-MS/MS' section) from 3 × 10-cm dishes of *Adnp*^{Flag-AviTag/Flag-AviTag} ES cells (cMB264). Nuclear lysates were then concentrated to 250 µl final volume using Amicon Ultra 0.5 ml Centrifugal Filters (3 kDa, Millipore), and fractionated by SEC on a Superose 6 HR 10/300 resin by fast protein liquid chromatography (AKTA; Amersham-Pharmacia Biotech). The predicted size exclusion maximum for this resin is 40 MDa, with a void volume of 7.35 ml. The column was equilibrated in 2 column volumes of gel filtration (GF) buffer (250 mM NaCl, 50 mM Tris-HCl pH 7.5, 1 mM DTT, 1 × PIC) before sample loading. A high-molecular-mass protein column standard was used to define the column resolution (Sigma). Protein peaks were detected by UV monitoring. Thyroglobulin (669,000 Da) peaked in fractions 9 and 10. Before loading, each nuclear lysate was adjusted to the appropriate column conditions and centrifuged at 100,000g for 30 min. A 200 µl of lysate was loaded onto the column and collected into 350-µl fractions; fractions were then subjected to trichloroacetic acid (TCA) precipitation for western blot analysis. For TCA precipitation, the sample volume was adjusted to 500 µl with the GF buffer followed by the addition of 50 µl 0.15% sodium deoxycholate; tubes were vortexed and incubated at room temperature for 10 min. Protein was precipitated by the addition of 25 µl of 100% TCA (Sigma), followed by a 20-min incubation at −20 °C. Precipitated proteins were collected by centrifugation at 10,000g for 10 min at 4 °C. Protein pellets were washed with acetone and air-dried. The protein pellet was solubilized in 1 × sample buffer (62.5 mM Tris, pH 6.8, 0.72 M β-mercaptoethanol or 0.1 M DTT, 10% glycerol, 2% SDS and 0.05% bromophenol blue) and resolved by NuPAGE-Novex Bis-Tris 4–12% gradient gels (Invitrogen) and subjected to western blot analysis (see 'Western blotting' section for further details).

In vitro biochemistry. For cloning, cDNA encoding full-length human ADNP (amino acid residues 1–1102) was PCR amplified with primers NotI-ADNP-forward (5'-AAAAAAGCGGCCGATGTTCCAACTTCTGTCAACAA-3') and KpnI-ADNP-reverse (5'-AAAAAAGGTACCCCTAGGCCTGTGGCTGCTC-3') and cloned into a pFast-Bac-derived vector (Invitrogen) in frame with an N-terminal His₆-tag. Plasmids encoding full-length or N-terminally truncated ADNP (amino acid residues 229–1102) with a C-terminal Strep-tag II were generated by PCR amplification of ADNP cDNA with primers NotI-ADNP-forward and KpnI-ADNP-C-reverse (5'-AAAAAAGGTACCGGCCCTGTTGGCTGCTCAGTT-3')

or NotI-ADNP(Δ N228)-forward (5'-AAAAAAGCGGCCGCATGCCAAAGTCCTATGAAGCTTT-3') and KpnI-ADNP-C-reverse. The amplified cDNA was cloned into a pAC8-derived vector³⁶. Expression constructs encoding full-length human HP1 γ (amino acid residues 1–183) were generated by amplification of cDNA using primers NotI-CBX3-forward (5'-AAAAAAGCGGCCGCATGGCCTCCAACAACTACATT-3') and KpnI-CBX3-reverse (5'-AAAAAAGGTACCTATTGAGCTTCATCTTCTGGA-3') and cloning into pFast-Bac-derived vectors in frame with an N-terminal His₆-tag or Strep-tag II. cDNA for individual chromodomain (amino acid residues 11–81) or CSD (amino acid residues 109–183) domains of HP1 γ was amplified with primers NotI-CBX3(11)-forward (5'-AAAAAAGCGGCCGCATGGGAAAAACAGAAATGGAAAG-3') and KpnI-CBX3(81)-reverse (5'-AAAAAAGGTACCTATTCTGAGAGTTAAGAAACGC-3') or NotI-CBX3(109)-forward (5'-AAAAAAGCGGCC GCGATGCTGCTGACAAACCAAGAG-3') and KpnI-CBX3-reverse and cloned into a pAC8-derived vector in frame with an N-terminal His₆-tag. cDNA encoding for full-length human CHD4 (amino acid residues 1–1912) was amplified with primers NotI-CHD4-forward (5'-AAAAAAGCGGCCGCATGGCCAGCGCCTGGGAT-3') and KpnI-CHD4-reverse (5'-AAAAAAGGTACCTACTGTTGCTGTGCAACCTG-3'). The resulting PCR product was cloned into a pAC8-derived vector in frame with an N-terminal His₆-tag.

In vitro reconstitution of ChAHP. Full-length and truncated versions of ChAHP subunits were subcloned into pAC8 or pFastBac-derived vectors³⁶. The following constructs were generated: human ADNP (amino acid residues 1–1102 or 229–1102) with a C-terminal Strep-tag II, N-terminally His₆-tagged human CHD4 (isoform 1, residues 1–1912) and N-terminally His₆-tagged variants of HP1 γ (residues 11–81 or 109–183) were cloned into pAC8-derived vectors. Full-length human ADNP (amino acid residues 1–1102) in frame with an N-terminal His₆-tag and full-length human HP1 γ (residues 1–183) in frame with an N-terminal Strep-tag II were cloned into pFastBac-derived vectors. Baculoviruses were generated in *Spodoptera frugiperda* Sf9 cells using the Bac-to-Bac method for pFastBac-derived vectors or by cotransfection with viral DNA for pAC8-based vectors. After one round of virus amplification in Sf9 cells, *Trichoplusia ni* High5 cells were infected with the respective Baculovirus (150 μ l of virus per 10 ml of High5 cells at a density of 2×10^6 cells ml⁻¹) and collected 48 h after infection. Cells were lysed by sonication in 50 mM Tris, pH 7.5, 300 mM NaCl, 5 mM β -mercaptoethanol, 0.1% Triton X-100, 1 mM PMSE, 1 \times PIC (Sigma-Aldrich). For pull-down experiments, cell lysate of a 15 ml culture was added to 30 μ l of Strep-Tactin Sepharose (IBA) or 30 μ l of His-tag purification resin (Roche) and incubated for 1 h at 4 °C. The beads were washed three times with lysis buffer, supplemented with 30 mM imidazole for histidine pull-down reactions. Proteins were eluted by addition of 2 \times sample buffer (62.5 mM Tris-HCl, pH 6.8, 2% SDS, 25% glycerol, 0.05% bromophenol blue and 5% β -mercaptoethanol) and analysed by SDS-PAGE and Coomassie staining.

For large-scale expression of the ChAHP complex, 1 l of Hi5 cells coinfectd with Baculoviruses encoding for His₆-tagged ADNP and Strep-tagged HP1 γ was combined with 2 l of Hi5 cells expressing His₆-tagged CHD4. Cells were lysed in lysis buffer and the cleared lysate was passed over a 50-ml Strep-Tactin Sepharose (IBA) column. The bound complex was eluted in 50 mM Tris-HCl, pH 7.5, 100 mM NaCl, 5 mM β -mercaptoethanol, 2.5 mM desthiobiotin, and bound to an anion-exchange chromatography column (Poros HQ) equilibrated in 50 mM Tris-HCl, pH 7.5, 100 mM NaCl, 5 mM β -mercaptoethanol. The bound proteins were eluted using a linear NaCl gradient, concentrated and further purified by SEC (HiLoad Superdex 200 26/600) in 50 mM HEPES-OH, pH 7.4, 150 mM NaCl and 0.5 mM TCEP. Fractions containing the ChAHP complex were concentrated and reinjected to a Superdex200 10/300 column equilibrated in the same buffer.

Computational methods. RNA-seq analysis. All sequencing reads were aligned to the December 2011 (mm10) mouse genome assembly from UCSC³⁷. HP1-mutant RNA-seq data were aligned using STAR 2.5.0a with the following settings to allow reporting of one randomly chosen alignment per multi-mapping read: '-outFilterMultimapNmax 20--outMultimapOrder Random--outSAMmultNmax 1--alignSJoverhangMin 8--alignSJDBoverhangMin 1--outFilterMismatchNmax 999--alignIntronMin 20--alignIntronMax 100000--alignMatesGapMax 100000--outSAMtype BAM SortedByCoordinate'. Aligned and sorted reads were indexed using SAMtools (version 1.2). ADNP-mutant RNA-seq data were aligned in Galaxy using Bowtie with the parameters '-m 1--best--strata'³⁸. Aligned Bam files were imported in R using QuasR (1.14.0)³⁹. BigWig files normalized for sequencing depth were generated using the QuasR qExportWig function. Reads were counted over exons using the qCount function and collapsed to yield one value per gene. This count table was used for differential expression calling with the EdgeR package⁴⁰. To compare the different *Cbx* knockout cell lines with *Adnp* knockouts (Extended Data Fig. 7), all biological replicates of the parental/untreated cell lines for *Cbx3* and *Adnp* were used as control group, whereas the respective knockout replicates were considered the treatment group.

Gene Ontology analysis. Gene Ontology term analysis of upregulated gene sets was performed using goana from the R limma package⁴¹. For the analysis significantly upregulated genes (FDR \leq 0.01, fold change \geq 4) from EdgeR output were used. **Repeat analysis.** RNA-seq libraries were mapped to the genome using STAR 2.5.0a with settings optimized for maximum repeat recovery/mappability (-outFilterType Normal--alignEndsType Local--winAnchorMultimapNmax 5000--seedPerWindowNmax 1000--alignTranscriptsPerReadNmax 100000--seedNoneLociPerWindow 100--alignWindowsPerReadNmax 20000--alignTranscriptsPerWindowNmax 1000--outFilterMultimapNmax 100000--outSAMattributes NH HI NM MD AS nM--outMultimapOrder Random--outSAMmultNmax 1). The resulting alignment file was intersected with repeat masker coordinates for mm10 (repeat masker 2012-02-07 update, downloaded from UCSC table browser), alignments overlapping repeats were counted for all repeat classes and normalized to 1 million mapping reads per library.

ChIP-seq read alignment. ChIP-seq data were aligned in R using the qAlign function from the QuasR package³⁹ with default settings, which calls the Bowtie aligner with parameters '-m 1--best--strata'³⁸. Depth-normalized BigWig files were generated using QuasR 1.14.0. For H3K9me3, STAR 2.5.0a (-alignIntronMax 1--alignEndsType EndToEnd--outFilterType Normal--seedSearchStartLmax 30--outFilterMultimapNmax 10000--outSAMattributes NH HI NM MD AS nM--outMultimapOrder Random--outSAMmultNmax 1--outSAMunmapped Within) was used and non-aligning and multiple mappers were filtered out using samtools. BigWig files displaying the full length for uniquely mapping reads were generated using bedtools and bedGraphToBigWig (UCSC binary utilities).

Peak finding. ADNP peaks were called on ChIP replicates using the corresponding inputs as background (all BAM files from QuasR alignment). MACS (version 2.1.1.20160309)⁴² was run with the default parameters. Peaks detected in at least two out of three replicates were kept.

HP1 γ peaks were called on both wild-type and knockout ADNP ChIP replicates individually, using the corresponding inputs as background (all BAM files from QuasR alignment). MACS was run with the following options: '-nomodel--shift 100--extsize 200. Subsequently, peak lists were intersected using bedtools intersect. Peaks present in both wild-type and knockout ADNP datasets, which did not contain the top scoring ADNP motif, were defined as ADNP-independent HP1 γ peaks. Note that the number of ADNP-independent HP1 γ peaks is an underestimate. HP1 proteins, particularly in heterochromatic regions, often cover large domains, similar to H3K9me3. However, for consistency reasons and to compile an accurate control group for comparison with the sharp ChAHP peaks, we chose to use MACS settings that are optimized for identification of transcription-factor like, narrow peaks rather than broad peaks. Hence the peaks called here represent a stringent set of the most highly enriched loci in the absence of ADNP. Most of the broader domains, which are also ADNP-independent, were not considered, as their shape is very different compared to ChAHP peaks.

Motif finding. HOMER v.4.8 was used with default settings to identify DNA sequence motifs in ADNP peaks⁴³.

Heat maps and meta-plots. Heat maps and meta-plots were generated from averaged replicates using the command line version of deepTools2⁴⁴. Peak centres were calculated based on the peak regions identified by MACS (see above). BigWig coverage files for individual replicates were generated by QuasR (see above). For averaging replicates and for calculating log₂(ChIP/input) ratios, bigwigCompare from deepTools2 was used. To generate histone modification meta-plots for ChAHP-bound loci (Extended Data Fig. 6), we used the following previously published datasets: H3K4me1 (GSE27841)⁴⁵, H3K4me2 and H3K27me3 (GSE25532)⁴⁶, H3K9ac (GSE31284)⁴⁷, H3K9me2 (GSE54412)⁴⁸, H3K9me3 (GSE12241)⁴⁹.

ATAC-seq analysis. Paired-end reads were aligned using STAR 2.5.0a using default parameters except for -alignIntronMax 1 and -alignEndsType EndToEnd. Only uniquely mapping reads (alignment score of 255) were kept for further analysis. These uniquely mapping reads were used to generate bigwig genome coverage files similar to ChIP-seq. Meta-profiles and heat maps were generated using deepTools2. For the meta-profiles, the average fragment count per 10-bp bin was normalized to the mean fragment count in the first and last five bins. This ensures that the background signal is set to one for all experiments.

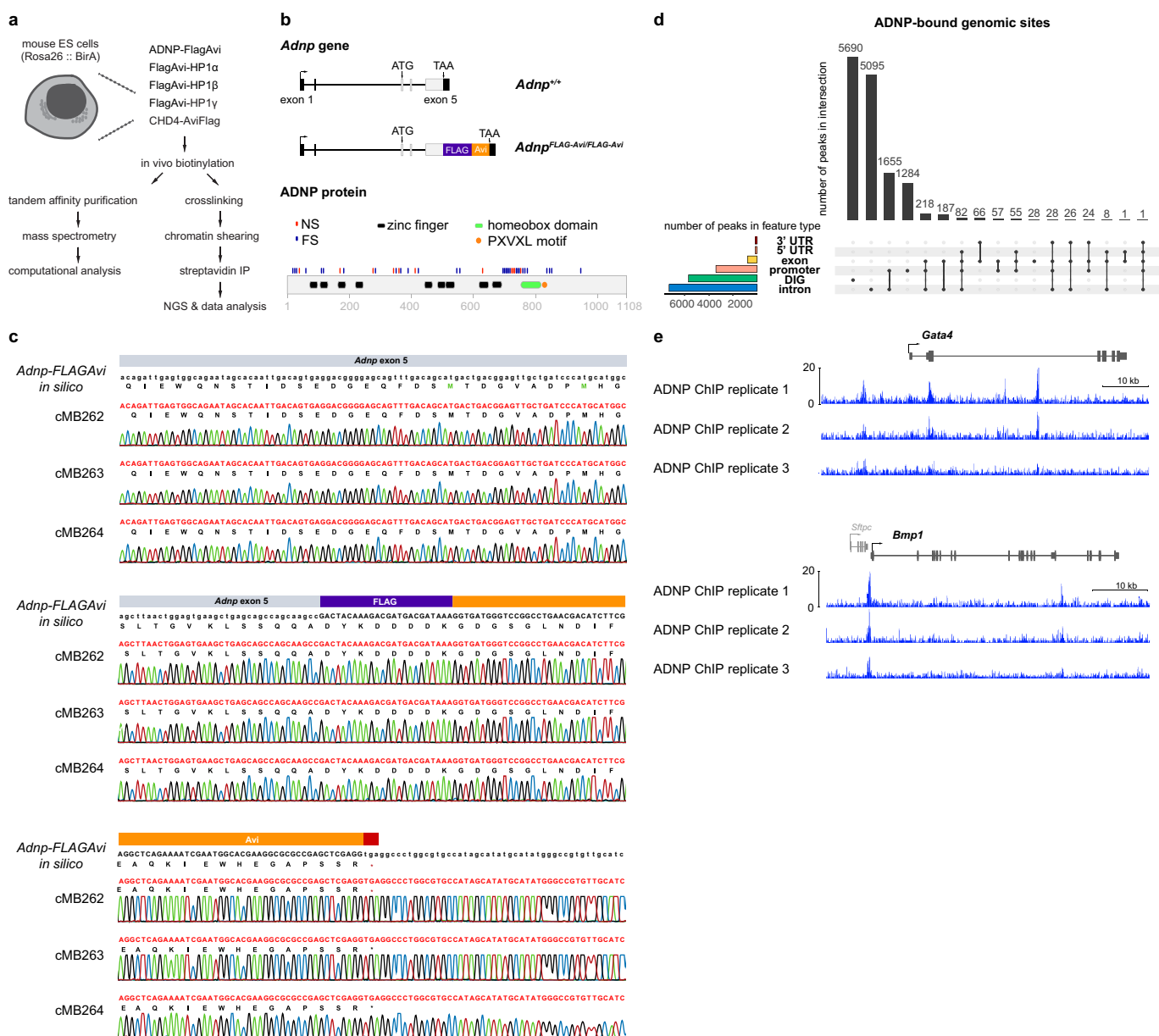
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. All custom codes used to analyse data and generate figures are available upon reasonable request.

Data availability. Genome-wide datasets are deposited at the Gene Expression Omnibus (GEO) under the accession number GSE97945. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE⁵⁰ partner repository with the dataset identifier PXD006226.

25. Mohn, F. et al. Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. *Mol. Cell* **30**, 755–766 (2008).

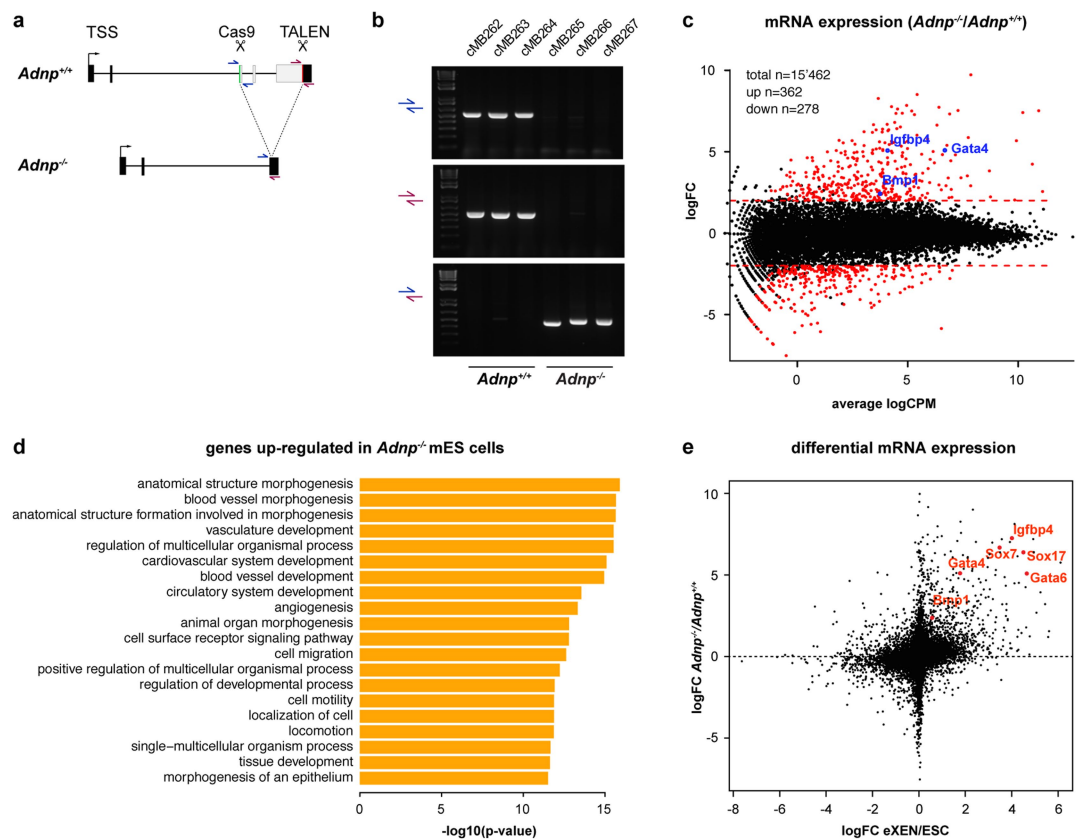
26. Knuckles, P. et al. RNA fate determination through cotranscriptional adenosine methylation and microprocessor binding. *Nat. Struct. Mol. Biol.* **24**, 561–569 (2017).
27. Jacobs, S. A. & Khorasanizadeh, S. Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. *Science* **295**, 2080–2083 (2002).
28. Bibel, M., Richter, J., Lacroix, E. & Barde, Y.-A. Generation of a defined and uniform population of CNS progenitors and neurons from mouse embryonic stem cells. *Nat. Protocols* **2**, 1034–1043 (2007).
29. Cox, J. et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
30. Cox, J. et al. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513–2526 (2014).
31. Hubner, N. C. et al. Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. *J. Cell Biol.* **189**, 739–754 (2010).
32. Tyanova, S. et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).
33. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* **98**, 5116–5121 (2001).
34. Schwanhäusser, B. et al. Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
35. MacLean, B. et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).
36. Abdulrahman, W. et al. A set of baculovirus transfer vectors for screening of affinity tags and parallel expression strategies. *Anal. Biochem.* **385**, 383–385 (2009).
37. Rosenbloom, K. R. et al. The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.* **43**, D670–D681 (2015).
38. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
39. Gaidatzis, D., Lerch, A., Hahne, F. & Stadler, M. B. QuasR: quantification and annotation of short reads in R. *Bioinformatics* **31**, 1130–1132 (2015).
40. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
41. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* **11**, R14 (2010).
42. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
43. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
44. Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44** (W1), W160–W165 (2016).
45. Whyte, W. A. et al. Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. *Nature* **482**, 221–225 (2012).
46. Tiwari, V. K. et al. A chromatin-modifying function of JNK during stem cell differentiation. *Nat. Genet.* **44**, 94–100 (2011).
47. Karimiyi, K., Krebs, A. R., Oulad-Abdelghani, M., Kimura, H. & Tora, L. H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. *BMC Genomics* **13**, 424 (2012).
48. Liu, N. et al. Recognition of H3K9 methylation by GLP is required for efficient establishment of H3K9 methylation, rapid target gene repression, and mouse viability. *Genes Dev.* **29**, 379–393 (2015).
49. Mikkelsen, T. S. et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
50. Vizcaino, J. A. et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44** (D1), D447–D456 (2016).
51. Kwon, S. H. & Workman, J. L. The heterochromatin protein 1 (HP1) family: put away a bias toward HP1. *Mol. Cells* **26**, 217–227 (2008).
52. Murzina, N., Verreault, A., Laue, E. & Stillman, B. Heterochromatin dynamics in mouse cells: interaction between chromatin assembly factor 1 and HP1 proteins. *Mol. Cell* **4**, 529–540 (1999).
53. Maison, C. et al. The SUMO protease SENP7 is a critical component to ensure HP1 enrichment at pericentric heterochromatin. *Nat. Struct. Mol. Biol.* **19**, 458–460 (2012).
54. Smothers, J. F. & Henikoff, S. The HP1 chromo shadow domain binds a consensus peptide pentamer. *Curr. Biol.* **10**, 27–30 (2000).



Extended Data Fig. 1 | Generation of isogenic mouse ES cell lines to interrogate protein–protein and protein–chromatin interactions.

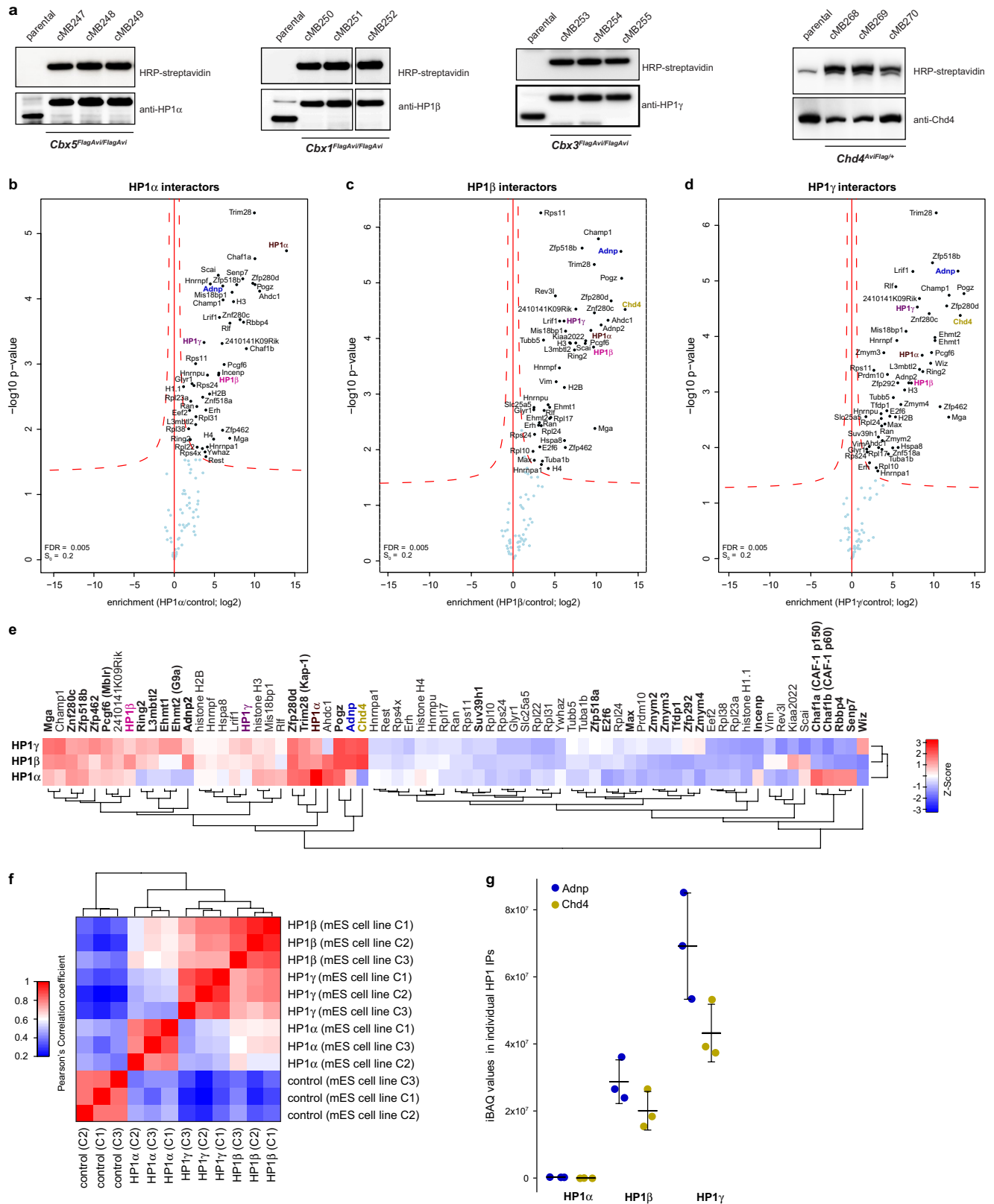
a, Mouse ES cells expressing the BirA biotin ligase from the *Rosa26* locus were used as a parental cell line for endogenous gene tagging with the Flag-AviTag¹². For a full list of mouse ES cell lines used in this study (cMB#), see Supplementary Information. **b**, Top, scheme depicting Flag-AviTag (not drawn to scale) insertion at the endogenous *Adnp* locus. Arrow indicates transcription start site. Boxes represent exons. Bottom, scheme depicting ADNP protein. Protein domains as predicted by InterPro. Nonsense (NS) and frameshift (FS) mutations found in children with Helsmoortel–Van der Aa syndrome are indicated (https://www.

adnpkids.com, dated March 2017). Numbers denote amino acids. **c**, Sanger sequencing of *Adnp*^{Flag-AviTag/Flag-AviTag} cell lines. **d**, Distribution of ADNP-bound genomic sites with respect to protein-coding genes. Peaks were called from ChIP-seq data acquired from three independent biological replicates (that is, three independent *Adnp*^{Flag-AviTag/Flag-AviTag} mouse ES cell lines). Horizontal bars represent peaks annotated to individual categories, and vertical bars represent peaks annotated jointly to specified combinations of categories. DIG, distal intergenic; UTR, untranslated region. **e**, ChIP-seq profiles at two lineage-specifying gene loci that were generated from three independent *Adnp*^{Flag-AviTag/Flag-AviTag} ES lines.



Extended Data Fig. 2 | Generation and analysis of isogenic *Adnp* knockout mouse ES cell lines. **a**, Scheme depicting CRISPR–Cas9 and TALEN-induced double-stranded DNA breaks to delete the *Adnp* open-reading frame. TSS, transcription start site. **b**, PCR genotyping confirming homozygous deletion of the *Adnp* open-reading frame in three different mouse ES cell lines used in this study. The experiment was performed twice. **c**, MA plot comparing fold change (FC) in gene expression for *Adnp*^{-/-} versus *Adnp*^{+/+} cells (y axis) with mean mRNA abundance

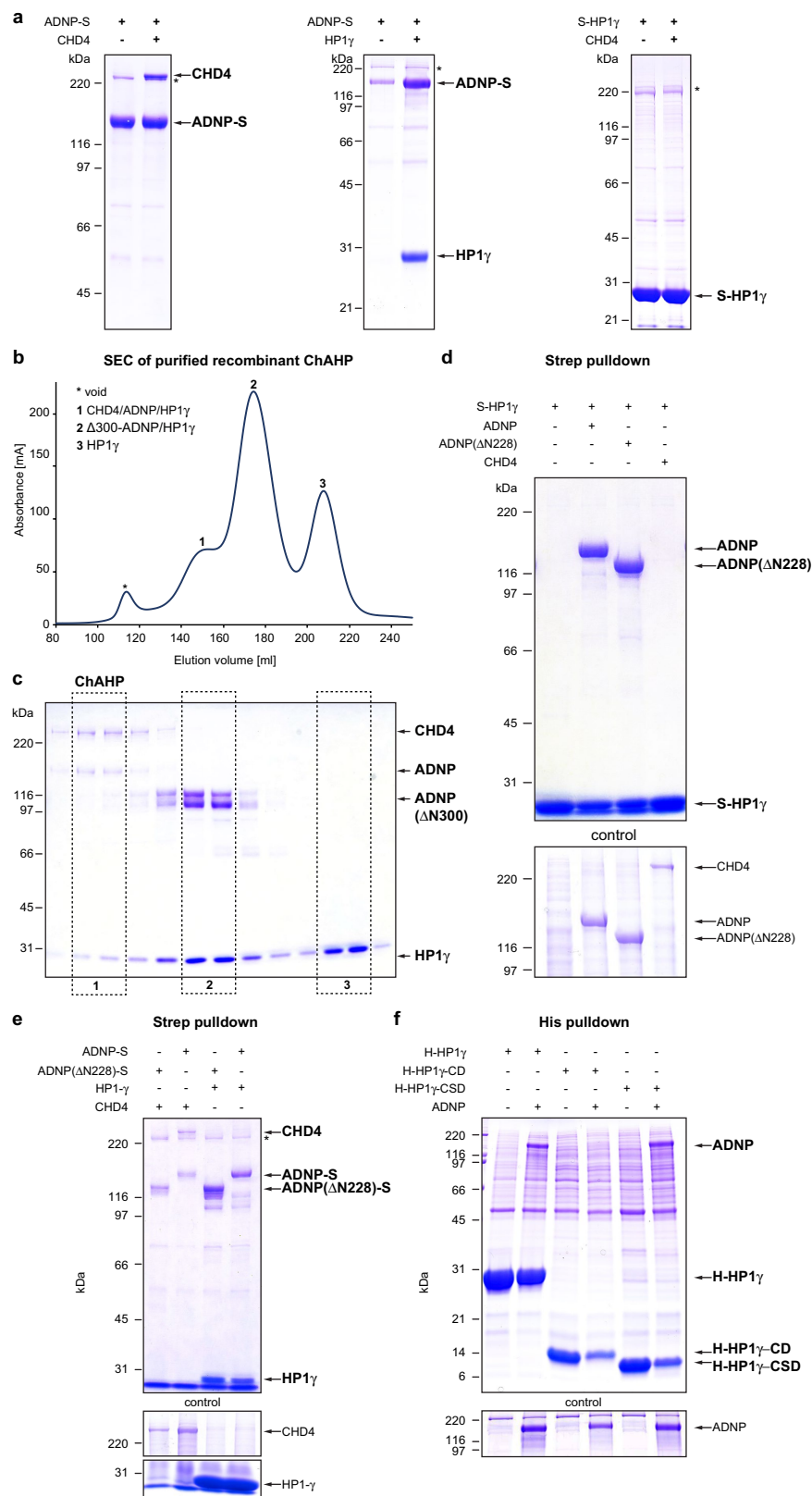
(x axis). Representative endoderm-specific genes are highlighted in red. Dashed red lines indicate fourfold up- or downregulation. CPM, counts per million. **d**, Gene Ontology enrichment analysis of genes upregulated in *Adnp*^{-/-} cells. *n* = 3 independent biological replicates. **e**, Scatterplot comparing gene expression fold change upon *Adnp* knockout (y axis) with expression changes between extraembryonic endoderm (eXEN) and ES cells (x axis). Known key lineage markers are indicated in blue.



Extended Data Fig. 3 | The HP1 interactome of mouse ES cells.

a, Isogenic mouse ES cell lines expressing endogenously tagged CHD4 and HP1 proteins. Western blot demonstrating expression of FlagAviTag-tagged proteins. The high molecular mass of CHD4 (218 kDa) does not allow discernable separation of tagged from non-tagged protein. See Supplementary Information for detailed genotype descriptions of the individual ES cell lines. For gel source data, see Supplementary Fig. 1. Experiments were performed twice. **b**, TAP-LC-MS/MS of endogenously FlagAviTag-tagged HP1 α . Protein purification was performed in the presence of 350 mM NaCl. Parental ES cell line serves as background control. $n = 3$ independent biological replicates (that is, three independent *Cbx5*^{Flag-AviTag/Flag-AviTag} mouse ES cell lines). **c**, TAP-LC-MS/MS of endogenously Flag-AviTag-tagged HP1 β . Protein purification was performed in the presence of 350 mM NaCl. Parental ES cell line serves as background control. $n = 3$ independent biological replicates (that is, three independent *Cbx1*^{Flag-AviTag/Flag-AviTag} mouse ES cell lines). **d**, TAP-LC-MS/MS of endogenously Flag-AviTag-tagged HP1 γ . Protein purification was performed in the presence of 350 mM NaCl. Parental mouse ES cell line serves as background control. $n = 3$ independent biological replicates

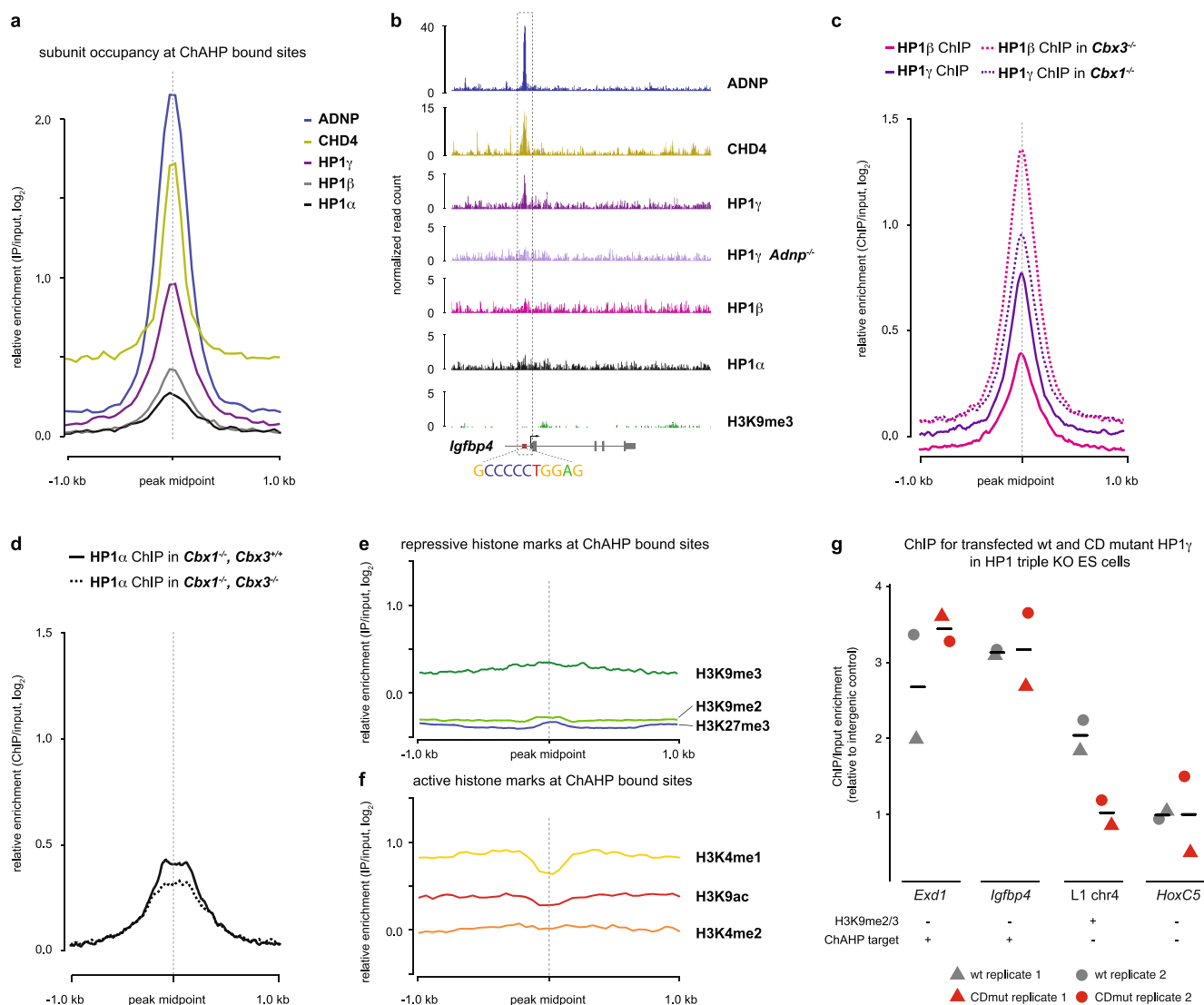
(that is, three independent *Cbx1*^{Flag-AviTag/Flag-AviTag} mouse ES cell lines). **e**, Heat map showing the variation in co-purifying (Z-score) proteins across HP1 isoform-specific TAP-LC-MS/MS experiments. Proteins that were significantly enriched in at least one experiment (**b–d**) were included in the analysis. **a–e**, Validating our approach, all three HP1 isoforms co-precipitated a large number of proteins. Many of these were common to all three HP1 proteins and have previously been described⁵¹. We also observed several proteins that interacted uniquely with specific isoforms, such as the previously identified CAF-1 or SENP7 interactions with HP1 α ^{52,53}. **f**, Heat map visualization of Pearson's correlation coefficients for the individual HP1 isoform-specific TAP-LC-MS/MS experiments. Three independent biological replicates for HP1 α and HP1 γ , two biological and one technical replicate for HP1 β , and three technical replicates for the parental cell line. **g**, iBAQ values of ADNP and CHD4 in HP1 isoform-specific TAP-LC-MS/MS experiments. Three independent biological replicates for HP1 α and HP1 γ , two biological and one technical replicate for HP1 β . Centre value denotes the mean; error bars denote s.d. **b–g**, Statistical analysis was performed using Perseus (see Methods). Mass spectrometry raw data are deposited with ProteomeXchange.



Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | In vitro characterization of ChAHP complex composition. **a**, Strep-tag pull-down assays with recombinant human proteins overexpressed in Hi5 insect cells, revealing that ADNP binds to both CHD4 and HP1 γ , whereas CHD4 and HP1 γ do not interact directly. **b**, SEC of the recombinant ChAHP complex. ChAHP was reconstituted from Hi5 insect cells and further purified by separation according to its molecular mass on a HiLoad Superdex 300 column. Largest fractions eluting first contain ChAHP (1), followed by ADNP-HP1 γ (2) and HP1 γ alone (3). **c**, Fractions from **b** were separated on SDS-PAGE and visualized by Coomassie staining. **d**, Pull-down analysis of Strep-tagged HP1 γ (S-HP1 γ) with full-length or N-terminally truncated ADNP (Δ N228) or CHD4. **e**, Pull-down analysis of Strep-tagged full-length or N-terminally truncated ADNP. **b–e**, Note that N-terminally truncated

ADNP does not co-elute with CHD4 on SEC (**b**, **c**). This is confirmed by pull-down experiments (**d**, **e**), which show that ADNP lacking the first 228 amino acids is only able to bind to HP1 γ but no longer to CHD4. Thus, ADNP contacts CHD4 through its N terminus. **f**, Pull-down analysis of His-tagged (H) full-length HP1 γ , and isolated chromodomain (CD) and chromoshadow domain (CSD). Similar to other proteins containing the conserved PXVXL pentapeptide⁵⁴, ADNP directly interacts with the CSD of HP1 γ . This is consistent with the previously reported interaction of ADNP with HP1 α ¹⁷. The chromodomain of HP1 γ does not bind to ADNP. Experiments in **a–f** were performed at least twice. S denotes the streptavidin tag added to the respective protein; asterisks denote a common contaminating protein in streptavidin pull-down assays. For gel source data, see Supplementary Fig. 1.



Extended Data Fig. 5 | HP1 occupancy at ChAHP-binding sites.

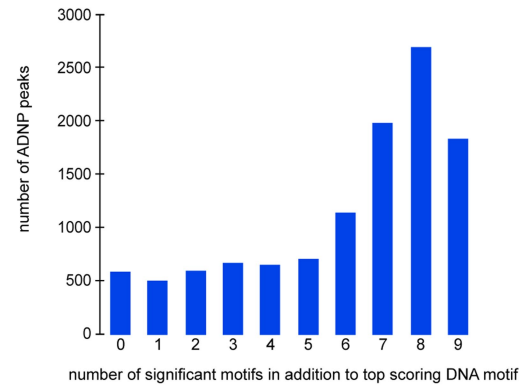
a, Subunit occupancy at ChAHP-bound sites displayed as meta-profile integrating signal of all peaks. **b**, Genome browser screen shot of the *Igf1bp4* locus. ChIP-seq tracks represent depth-normalized read counts of averaged replicate experiments. The predicted ADNP DNA-binding motif upstream of the *Igf1bp4* transcription start site is shown. **c**, Average HP1 β and HP1 γ ChIP-seq enrichment on ChAHP-bound sites in wild-type cells, and average HP1 β and HP1 γ ChIP-seq enrichment on ChAHP-bound sites in *Cbx3*^{-/-} and *Cbx1*^{-/-} mouse ES cell lines, respectively. *n* = 2 biological replicates (that is, independently tagged mouse ES cell lines). **d**, Average HP1 α ChIP-seq enrichment on ChAHP-bound sites in wild-type and *Cbx1*^{-/-}*Cbx3*^{-/-} double-knockout ES cell lines. *n* = 2 biological replicates. **e**, Histone modifications associated with heterochromatin are absent at ChAHP-bound sites. **f**, Histone modifications associated

with active transcription are absent at ChAHP-bound sites. **e**, **f**, Histone modification profiles are displayed as meta-profile integrating signal over all peaks. **g**, Binding of wild-type and chromodomain mutant HP1 γ to ChAHP targets (*Igf1bp4* and *Exd1*), an H3K9me3-modified region next to an L1 repeat (L1 chr4) and an inactive promoter of an unrelated gene (*HoxC5*), quantified by ChIP-qPCR. Fold enrichment was normalized to an intergenic control region devoid of HP1 γ and H3K9me3. Wild-type (grey) and mutant (red) HP1 γ constructs were transiently transfected into HP1 triple-knockout ES cells in biological duplicates. Note the decrease of HP1 γ binding at the H3K9me3-modified region (L1 chr4), whereas ChAHP targets remain unaffected in the chromodomain mutant (CDmut) that can no longer bind to H3K9me3. Black lines indicate average enrichments.

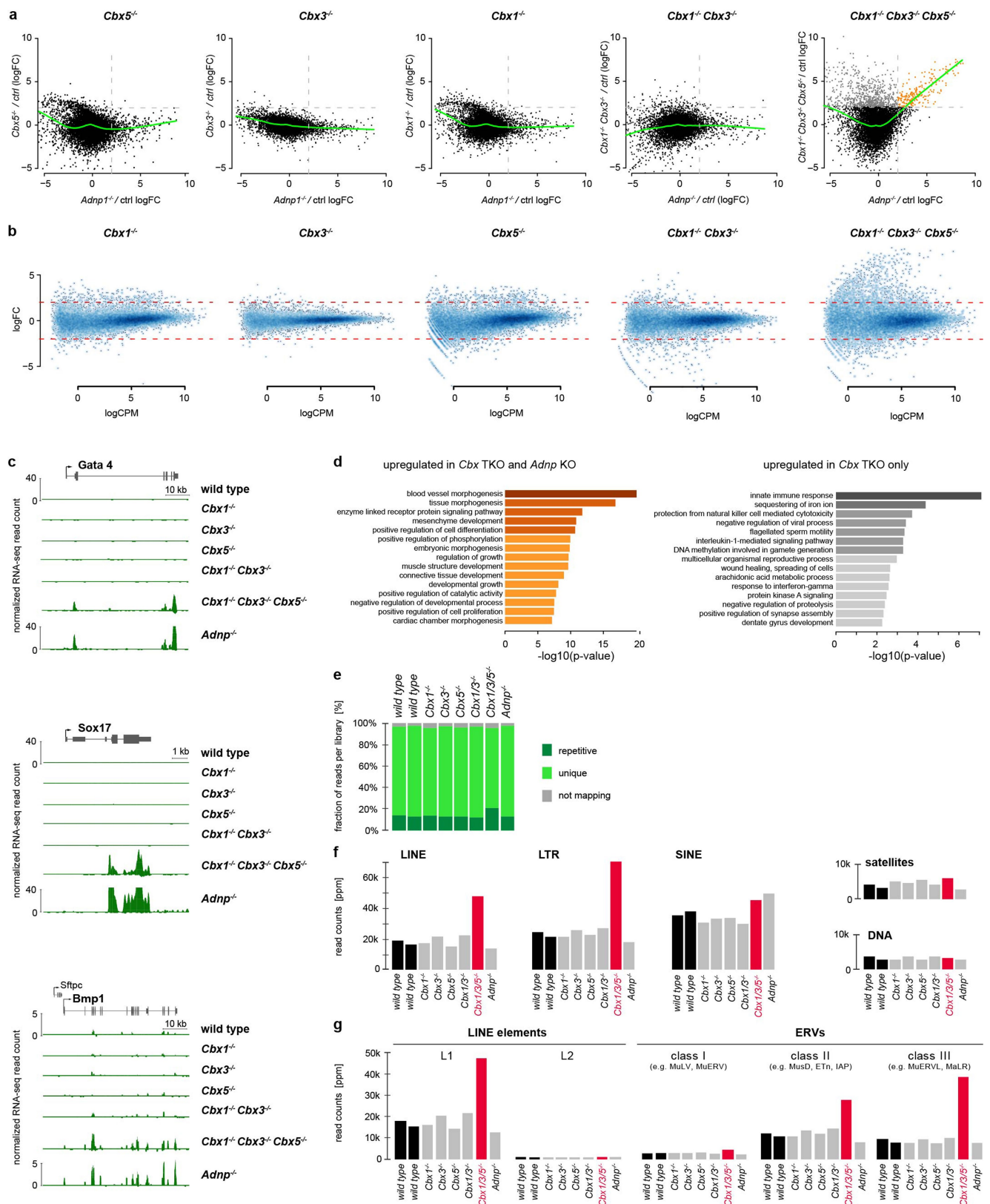
a**Table of significant motifs in ADNP peaks**

Rank	Motif	P-value	% of Targets	% of Background
1	CGCCCCCTTCTG	1e-10538	63.22%	2.80%
2	GTTCAATTCCTA	1e-6714	55.28%	4.66%
3	ATCTGGCGCCCT	1e-6478	34.95%	0.94%
4	CCACCCATATGG	1e-6104	55.26%	5.56%
5	GGTTATGAGC	1e-5759	56.24%	6.49%
6	TAAGAGCACTCA	1e-5416	49.14%	4.75%
7	AGATGGCTCA	1e-5379	52.92%	6.03%
8	CAGACACACC	1e-5226	29.68%	0.89%
9	GGGTCTCTGG	1e-4742	42.68%	3.89%
10	GTAATGAGATCT	1e-4374	36.99%	2.90%
11	GTCTTAAGACAG	1e-4305	33.13%	2.10%
12	GGACCTCTGG	1e-3839	37.99%	3.96%
13	ACAACCCAC	1e-3770	46.29%	7.04%
14	CTACAGTGTACT	1e-3626	29.57%	2.06%
15	TTCGGAG	1e-2262	37.57%	7.74%
16	TATATAAATAAA	1e-1994	23.86%	3.07%
17	CCTGGAGA	1e-1804	30.53%	6.25%
18	ATTTATTI	1e-1560	26.67%	5.41%
19	TCCGAGCC	1e-1341	26.54%	6.24%
20	TTTTTTTAAAAA	1e-853	19.09%	4.85%
21	GGCACTGCACCT	1e-471	6.50%	0.94%
22	GTAACCTCAGTT	1e-414	9.58%	2.41%
23	TGTCCTCGAACT	1e-98	7.18%	3.57%

Extended Data Fig. 6 | Motif analysis of ADNP-bound loci. a, ADNP DNA-binding motifs predicted by HOMER. Frequency of occurrence and *P* values for motif enrichment compared to genomic background are indicated. *n* = 3 independent cell lines. **b**, Analysis of co-occurrence of the top-ten scoring ADNP DNA motifs. The bar graph shows the

b**Co-occurrence of top 10 motifs in peaks containing the top scoring motif.**

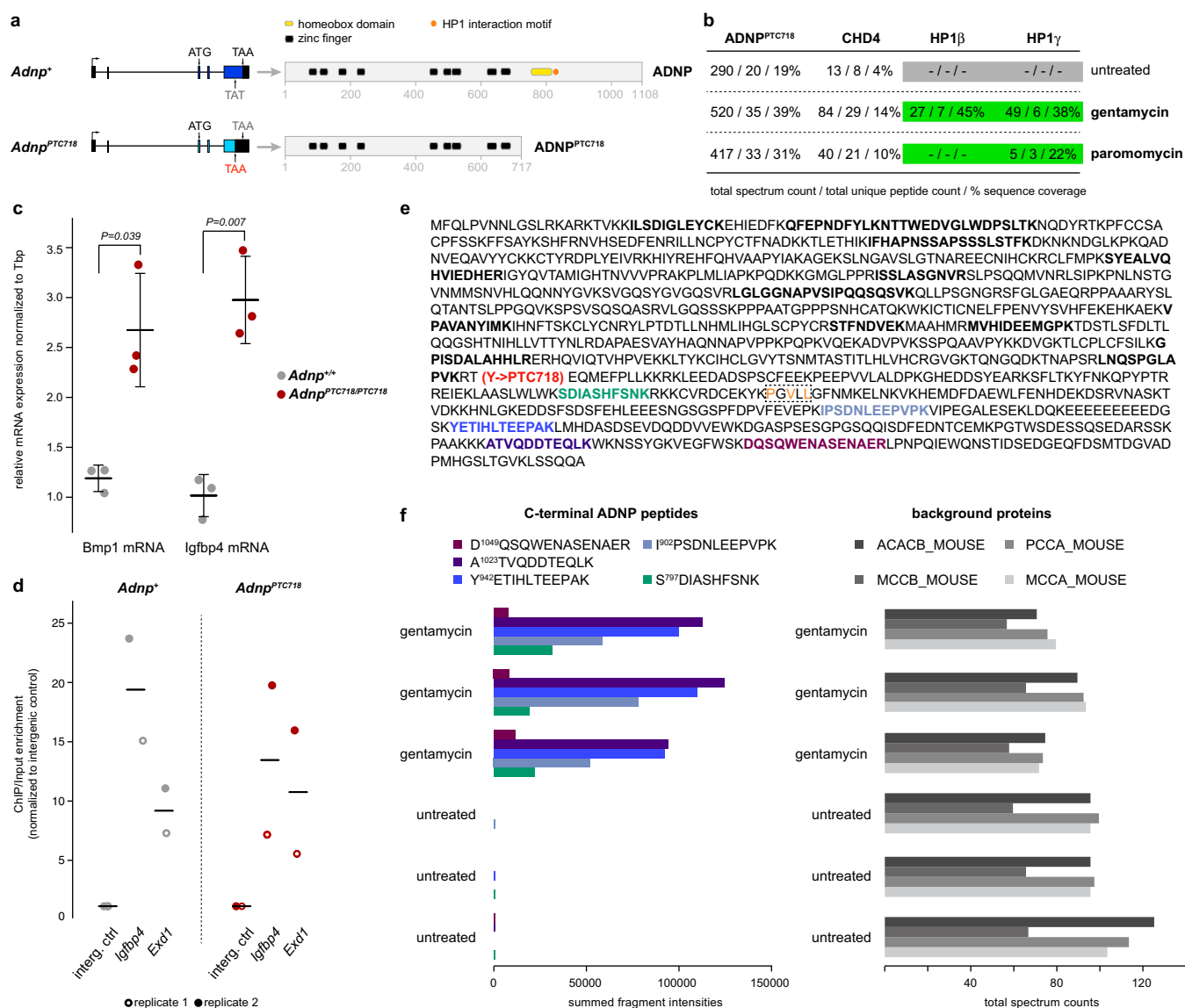
frequency of peaks containing the top-scoring ADNP motif and up to nine additional motifs as indicated on the x axis. Note that most peaks besides the GCCCCCTGGAG motif also contain more than five other sequence motifs out of the top-ten list.



Extended Data Fig. 7 | See next page for caption.

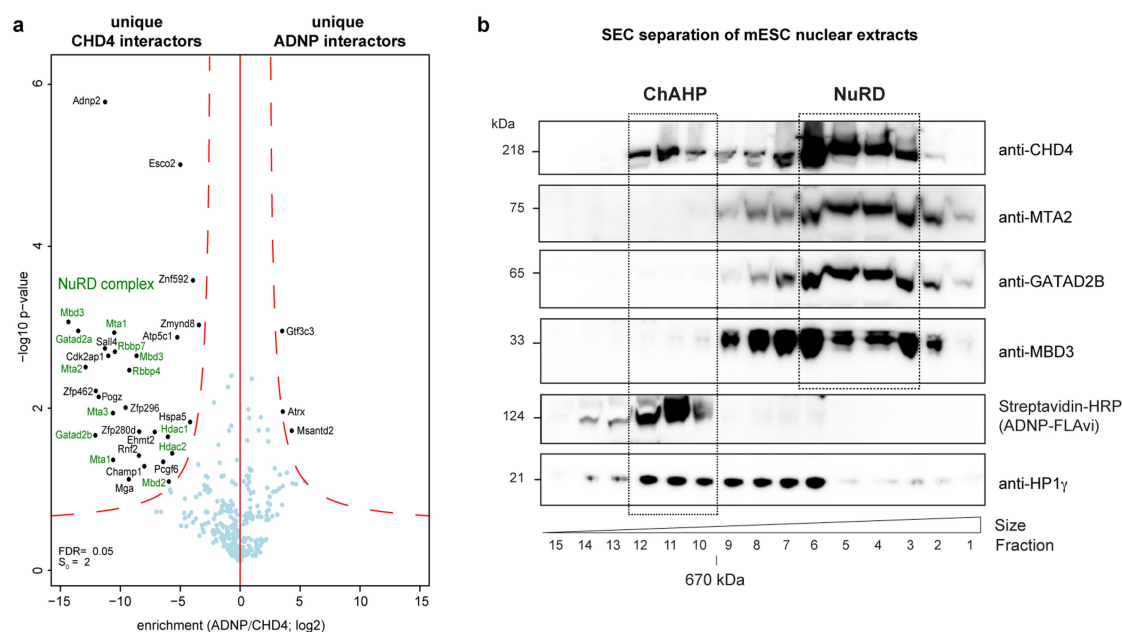
Extended Data Fig. 7 | Different HP1 isoforms can functionally substitute each other. **a**, Scatterplots comparing mRNA expression changes after deletion of *Adnp* versus single, double or triple deletions of *Cbx* genes measured by RNA-seq. Green trend lines indicate a loess (locally weighted scatterplot smoothing) regression. $n = 3$ biological replicates (that is, three independent *Adnp*^{-/-}, *Cbx1*^{-/-}, *Cbx3*^{-/-}, *Cbx5*^{-/-}, *Cbx1*^{-/-}*Cbx3*^{-/-} double knockout, or *Cbx1*^{-/-}*Cbx3*^{-/-}*Cbx5*^{-/-} triple knockout mouse ES cell lines). **b**, MA plot displaying fold changes in gene expression for individual *Cbx* knockout cell lines versus wild type. x axis denotes the mean mRNA abundance, $\log_2(\text{counts per million})$; y axis denotes the $\log_2(\text{fold change})$ between knockout and wild type. Dashed red lines indicate fourfold up- or downregulation. **c**, UCSC genome browser shots of three lineage-specifying genes. RNA-seq profiles normalized by library size of representative wild-type and

mutant ES cell lines are shown. Experiments were performed three times. **d**, Gene Ontology enrichment analysis of the genes that are upregulated in *Cbx1*^{-/-}*Cbx3*^{-/-}*Cbx5*^{-/-} triple-knockout (TKO) and *Adnp*^{-/-} knockout (KO) cells (orange group of genes in **a**), and of the genes that are upregulated in *Cbx1*^{-/-}*Cbx3*^{-/-}*Cbx5*^{-/-} triple-knockout but not *Adnp*^{-/-} knockout cells (grey group of genes in **a**). See also Supplementary Table 5. $n = 3$ independent cell lines. **e**, RNA-seq library statistics showing fraction of uniquely, multi- and non-mapping reads. Note the increase in multi-mappers in the HP1 triple-knockout cells. **f**, Quantification of reads mapping to the major repeat classes in counts per million mappable reads. **g**, Quantification of reads mapping to the different LINE and LTR elements in counts per million mappable reads. All mutant cell lines were derived from the same parental mouse ES cell line through direct genome editing and are therefore isogenic.



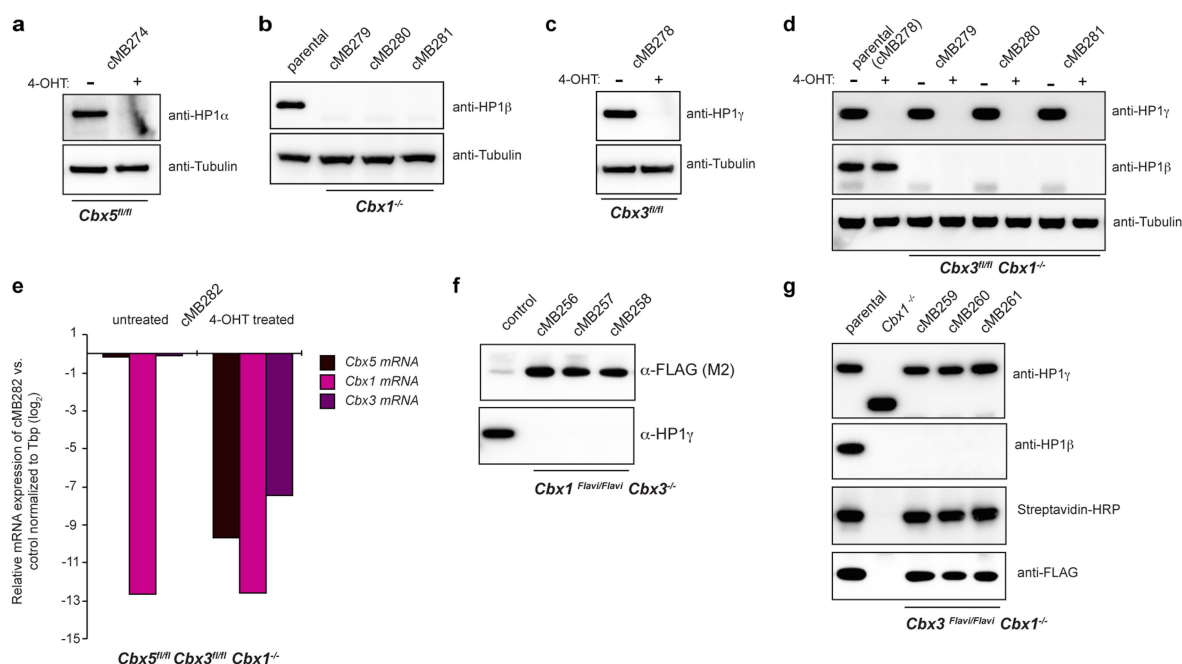
Extended Data Fig. 8 | A patient-specific nonsense mutation in *Adnp* impairs the interaction with HP1 but not with DNA. **a**, Scheme depicting the wild-type and mutant *Adnp* alleles, which code for Tyr (blue) and a patient-specific premature termination codon (red) at amino acid position 718, respectively. Full-length and truncated protein products are shown on the right. Arrow indicates transcription start site. Boxes represent exons. Numbers denote amino acids. **b**, N-terminally Flag-AviTag-tagged ADNP^{PTC718} was streptavidin-purified from cells with and without aminoglycoside treatment (gentamycin or paromomycin) and subjected to LC-MS/MS analysis. ADNP^{PTC718}-expressing cells were treated with 2 mg ml⁻¹ gentamycin (2.9 mM) or paromomycin (3.2 mM) for 24 h. The table depicts total spectral counts, unique peptides and percentage sequence coverage (derived from Scaffold) for all ChAHP components from the different treatments. **c**, qRT-PCR measurement of *Bmp1* and *Igfbp4* mRNA levels in ES cells expressing full-length *Adnp* (*Adnp*^{+/+}) or C-terminally truncated *Adnp* that interacts with CHD4 but

not with HP1 (*Adnp*^{PTC718/PTC718}). *n* = 3 biological replicates (that is, three independent RNA isolations). *P* values were calculated using two-tailed unpaired unequal variances *t*-tests. Centre value denotes the mean; error bars denote s.d. **d**, ChIP-qPCR enrichments for transiently transfected Flag-AviTag-tagged wild-type ADNP and ADNP^{PTC718} constructs on two ADNP targets, normalized to an intergenic control. Black lines indicate means. **e**, C-terminally Flag-AviTag-tagged ADNP^{PTC718} was streptavidin-purified from cells with or without gentamycin treatment (2.9 mM) and subjected to LC-MS/MS analysis. Bold letters indicate unique peptides further quantified by parallel reaction monitoring (PRM). C-terminal peptides encoded downstream of PTC718 are shown in colour. Dashed box denotes the HP1 interaction motif. **f**, Summed fragment intensities of five C-terminal ADNP peptides that are encoded downstream of PTC718 are shown on the left. Background proteins shown on the right serve as loading controls. Intensities were measured by PRM. Total spectrum counts were derived from Scaffold. *n* = 3 biological replicates.



Extended Data Fig. 9 | ChAHP and NuRD are distinct protein complexes. **a**, Single-step purification followed by LC-MS/MS of endogenously Flag-AviTag-tagged CHD4 and ADNP. Protein purification was performed in the presence of 350 mM NaCl. Proteins that interact predominantly with CHD4 or ADNP are indicated by UniProt names. NuRD complex components are labelled in green. $n = 3$ biological replicates (that is, three independent *Chd4*^{Flag-AviTag/Flag-AviTag} and

Adnp^{Flag-AviTag/Flag-AviTag} ES cell lines). Statistical analysis was done with Perseus (see Methods for details). Mass spectrometry raw data are deposited with ProteomeXchange. **b**, SEC of nuclear protein extracts from *Adnp*^{Flag-AviTag/Flag-AviTag} ES cells. Each fraction (indicated at the bottom) was resolved by SDS-PAGE and immunoblotted with the indicated antibodies. Molecular mass of individual proteins is indicated on the left. For gel source data, see Supplementary Fig. 1. Experiment was performed twice.



Extended Data Fig. 10 | Isogenic *Cbx* knockout ES cell lines. **a**, Western blot demonstrating depletion of HP1α protein in *Cbx5*^{fl/fl} mouse ES cell line after treatment with 4-hydroxytamoxifen (4-OHT). **b**, Western blot demonstrating depletion of HP1β protein in three independent *Cbx1*^{-/-} ES cell lines. **c**, Western blot demonstrating depletion of HP1γ protein in *Cbx3*^{fl/fl} ES cell line after treatment with 4-OHT. **d**, Western blot demonstrating depletion of HP1β and HP1γ proteins in three independent *Cbx1*^{-/-} *Cbx3*^{fl/fl} double-knockout cell lines after treatment with 4-OHT.

n = 3 independent 4-OHT treatments. **e**, qRT-PCR demonstrating depletion of *Cbx5*, *Cbx1* and *Cbx3* mRNAs in the *Cbx1*^{-/-} *Cbx3*^{fl/fl} *Cbx5*^{fl/fl} triple-knockout cell line upon treatment with 4-OHT. **f**, Western blot demonstrating depletion of HP1γ protein in three independent *Cbx1*^{Flag-Flag} *Cbx3*^{-/-} cell lines. **g**, Western blot demonstrating depletion of HP1β protein in three independent *Cbx3*^{Flag-Flag} *Cbx1*^{-/-} cell lines. For gel source data shown, see Supplementary Fig. 1. Experiments were performed twice.

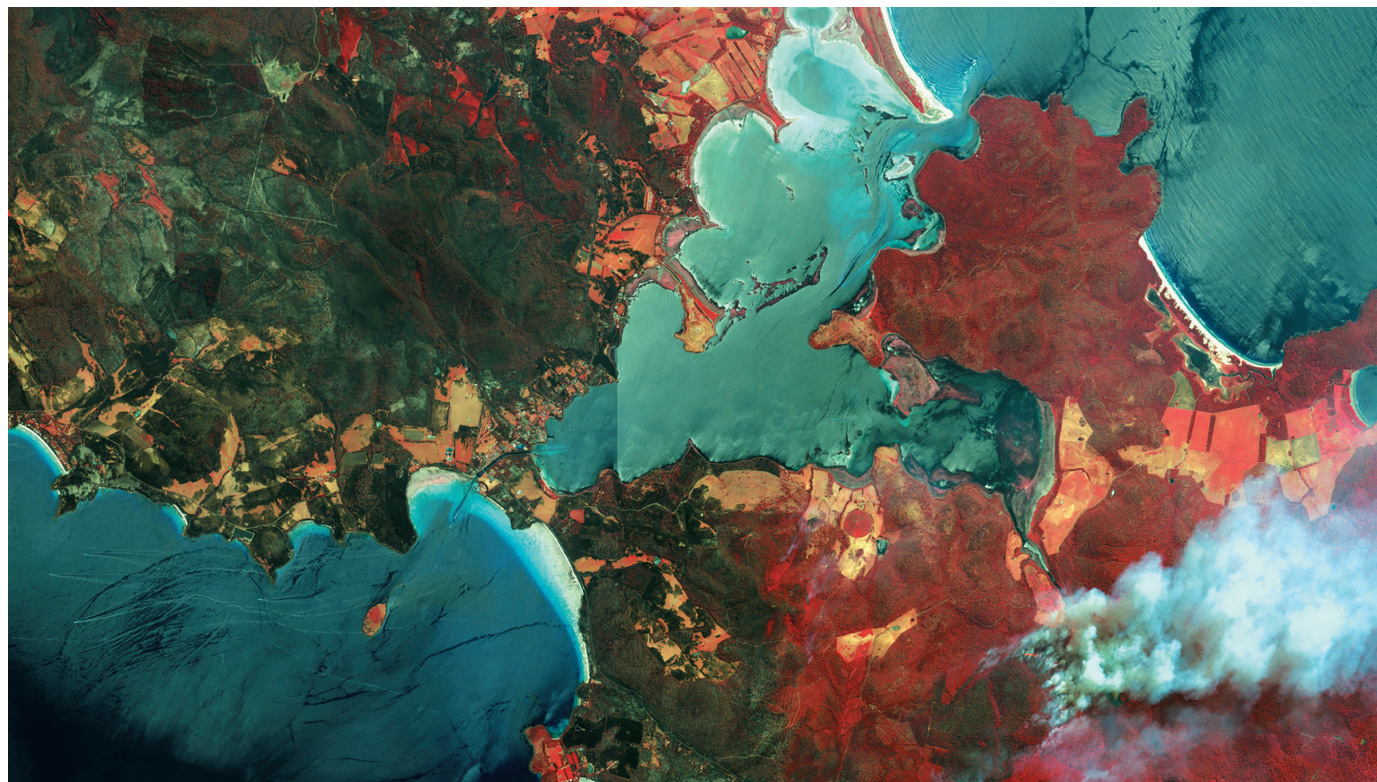
CAREERS

PUBLISHING Men expected to outnumber women as authors for next 80 years **p.747**

CONDUCT Petitioners urge US science academy to drop harassers **p.747**

NATUREJOBS For the latest career listings and advice www.naturejobs.com

DIGITALGLOBE/GETTY



Satellite images such as this one, depicting wildfires in Tasmania, Australia, can be used to track and visualize threats to the natural environment.

IMAGERY

Crunch time for data

Satellite and cloud-computing services are opening worlds of opportunity for researchers.

BY GABRIEL POPKIN

Samapriya Roy remembers when it would take him up to an hour to download a single 1-gigabyte image taken by the Landsat Earth-imaging satellites. That was in the late 2000s, when he was analysing satellite imagery as part of his undergraduate studies at Visvesvaraya National Institute of Technology in Maharashtra state, India. And the computer analysis of a picture could take even longer. Sometimes Roy would start the analysis at night and it would still be running the next morning.

Things are very different nowadays. Roy, who is a PhD student at Indiana University in Bloomington, uses a Google platform to store his data and run his algorithms and is able to crunch tens of thousands of images in minutes; all he needs is a web browser. “It brings everyone

to a level playing field,” he says. In addition to data from US government sources such as Landsat, he uses sharp, detailed images from three commercial satellite companies — two of which didn’t exist when he was an undergraduate — to research coastal land loss in Louisiana and the Amazon region of Brazil.

In the past few years, technology and satellite companies’ offerings to scientists have increased dramatically. Thousands of researchers now use high-resolution data from commercial satellites for their work. Thousands more use cloud-computing resources provided by big Internet companies to crunch data sets that would overwhelm most university computing clusters. Researchers use the new capabilities to track and visualize forest and coral-reef loss; monitor farm crops to boost yields; and predict glacier melt and

disease outbreaks. Often, they are analysing much larger areas than has ever been possible — sometimes even encompassing the entire globe. Such studies are landing in leading journals and grabbing media attention.

Commercial data and cloud computing are not panaceas for all research questions. NASA and the European Space Agency carefully calibrate the spectral quality of their imagers and test them with particular types of scientific analysis in mind, whereas the aim of many commercial satellites is to take good-quality, high-resolution pictures for governments and private customers. And no company can compete with Landsat’s free, publicly available, 46-year archive of images of Earth’s surface. For commercial data, scientists must often request images of specific regions taken at specific times, and agree not to publish raw data. Some ►

► companies reserve cloud-computing assets for researchers with aligned interests such as artificial intelligence or geospatial-data analysis. And although companies publicly make some funding and other resources available for scientists, getting access to commercial data and resources often requires personal connections. Still, by choosing the right data sources and partners, scientists can explore new approaches to research problems.

MAPPING POVERTY

Joshua Blumenstock, an information scientist at the University of California, Berkeley (UCB), is always on the hunt for data he can use to map wealth and poverty, especially in countries that do not conduct regular censuses. “If you’re trying to design policy or do anything to improve living conditions, you generally need data to figure out where to go, to figure out who to help, even to figure out if the things you’re doing are making a difference.”

In a 2015 study, he used records from mobile-phone companies to map Rwanda’s wealth distribution (J. Blumenstock *et al. Science* **350**, 1073–1076; 2015). But to track wealth distribution worldwide, patching together data-sharing agreements with hundreds of these companies would have been impractical. Another potential information source — high-resolution commercial satellite imagery — could have cost him upwards of US\$10,000 for data from just one country.

Blumenstock then learnt that Facebook had bought commercial satellite images for a programme it launched in 2014 to connect the global population to the Internet. After chats with a Facebook researcher on the project, he and the social-networking giant hammered out an agreement. Facebook would fund one of his graduate students to use the company’s technology to study how economic data from public surveys correlated with the visual characteristics of buildings represented in the satellite data. Facebook, in turn, could potentially gain a sharper view of the socio-economic characteristics of rural areas, whose residents are least likely to have Internet connections. (Facebook declined to comment.)

The arrangement presented some challenges, however. Facebook demanded a non-disclosure agreement before sharing data. (Blumenstock does not have access to personal Facebook user data, only to satellite and other aggregated data.) And UCB industry-partnership specialists scrutinized the agreement to ensure that it wouldn’t compromise academic integrity. Privacy concerns are likely to loom larger from now on. In the wake of allegations in March that a UK consultancy had deployed Facebook user data for US political purposes, universities and companies might be examining their agreements more closely.

Facebook’s command of machine learning and cloud computing was also the main draw for Robert Chen, a geographer at Columbia University in New York City, who collaborates



CONNIE ZHOU/GOOGLE/ZUMA PRESS/EYEVIEW

Google’s data centres support its cloud-based Earth Engine platform, which processes geospatial data.

with the company to study global population distribution. Data crunching that would have once taken years was completed in weeks, enabling Chen and his colleagues to produce high-resolution population maps of rural areas in 18 countries around the world (see go.nature.com/2s1dgg4). “Facebook can process 14.5 billion images in a couple of weeks,” he says. The social-media firm’s main goal for the project is to provide global Internet access (and reach more potential users). Chen aims to apply the maps to humanitarian assistance, conservation and development planning.

Other hi-tech goliaths are making resources available to researchers. Microsoft’s AI for Earth, which launched in late 2017, has enabled more than 60 research groups from more than 20 countries to analyse remote-sensing data sets from Esri, a mapping and geospatial-analysis company in Redlands, California, using Microsoft’s artificial intelligence (AI) algorithms and computing power. Microsoft’s chief environmental scientist, Lucas Joppa, says that AI can supercharge remote-sensing research by ferreting out previously hidden patterns in data. For example, a team including Milind Tambe, a computer scientist at the University of Southern California in Los Angeles, has used Microsoft algorithms to predict wildlife-poaching activity in Africa from drone imagery (see go.nature.com/2s2z5ta).

Researchers apply online for initial access to the program. If Joppa and his colleagues find a project promising, they collaborate and share expertise and in-kind resources, such as computing time, to help the research advance.

Amazon Web Services, the cloud-computing

branch of the e-commerce giant Amazon, started hosting the Landsat archive in early 2015. In September 2016, the company launched its Earth on AWS programme, through which it hosts around 15 data sets, including imagery, weather data from the US National Oceanic and Atmospheric Administration, and air-quality data from the non-profit organization OpenAQ in Washington DC. Although anyone can pay to analyse the data using Amazon’s computers, scientists can apply for donations of computing time; applications must include a description of the research problem and plans for dissemination of the results.

Google now hosts more than 600 public satellite, weather, population and other Earth and environmental data sets through its Earth Engine platform. More than 70,000 users — most of them researchers — have created free accounts on the platform, says Rebecca Moore, Earth Engine’s director of engineering.

The first global study done on the platform yielded a blockbuster paper on maps of forest change based on Landsat data; it has racked up nearly 3,000 citations in less than 5 years (M. C. Hansen *et al. Science* **342**, 850–853; 2013). Google’s infrastructure jump-started the project in 2013 by turning what would have been 15 years’ worth of data crunching on one computer into a job that took just a few days, says Matthew Hansen, a geographer at the University of Maryland in College Park who led the study.

The platform has since supported global studies of surface water, fish stocks, urban agriculture and transport networks, as well as smaller-scale studies. For Daniel Weiss, an epidemiologist at the University of Oxford, UK, who used Earth Engine to map travel time from any point on the globe to the nearest city (see go.nature.com/2ibwhbm), the platform

“Around 20 companies worldwide now offer or plan to offer Earth-observing capabilities.”

efficiently crunched a computationally expensive algorithm, saving months of work. The map itself is now a public resource on Earth Engine, and Weiss and his team are using it to produce better forecasts of malaria outbreaks.

MORE THAN PRETTY PICTURES

The growing fleet of satellite companies is serving up an increasingly diverse menu of data and images. Around 20 companies worldwide now offer or plan to offer Earth-observing capabilities. These firms, which have conventionally served military and private-sector clients in finance, agriculture and other arenas, are increasing their overtures to scientists.

In 2017, satellite company DigitalGlobe in Boulder, Colorado, provided scientists with high-resolution images worth around \$6 million through its DigitalGlobe Foundation, according to the foundation's president, Kumar Navulur. For some researchers, the company's super-sharp satellite-borne cameras have enabled previously difficult or impossible studies. Sarah Parcak, for example, an archaeologist at the University of Alabama at Birmingham, has used DigitalGlobe imagery to discover hidden sites in Egypt and elsewhere, and to track looting incidents.

Satellogic, a company in Buenos Aires founded in 2010, has promised to make hyperspectral data — information-rich imagery derived from light in dozens of wavelength bands — available to any scientist who wants them. No public satellite currently collects such data, which many scientists prize for its usefulness in applications such as detecting drought stress in plants and exploring for minerals. The company says that it has shared hyperspectral data with around two dozen researchers; Roy says he got access to some data for his Louisiana research after an e-mail exchange.

The satellite company Planet, based in San Francisco, California, images the globe daily, the side of each pixel in an image representing between 3 and 5 metres on the ground. The company makes data available to scientists through its research and education programme, which offers free data for up to 10,000 square kilometres a month to scientists who apply.

Institutions can also take out subscriptions for larger data volumes. Planet has provided imagery to more than 1,600 researchers from more than 70 countries, according to Joseph Mascaro, the company's director of academic programmes. The company's frequent images enabled Andreas Kääb, a geoscientist at the University of Oslo, to track melting glaciers in near-real time in Tibet, which showed that weather and climate change caused the glaciers to suddenly collapse (A. Kääb *et al.* *Nature Geosci.* **11**, 114–120; 2018). In 2016, he had warned the Chinese government of

an impending avalanche in Tibet on the basis of signals he had detected in Planet's images.

Kääb's research has benefited not just from the imagery itself but also from access to company staff, he says. "We typically write to Joe [Mascaro] and he connects us to someone from the team," Kääb says. "I feel to some extent I am part of the game, part of the process."

Using commercial data can have downsides. Companies such as DigitalGlobe and Satellogic typically take pictures that paying customers request, so scientists might find that no data are available for their area or time of interest. Government restrictions can also limit data availability. Mascaro and Navulur are prohibited by US law from sharing extremely high-resolution imagery of certain countries such as Israel, and cannot share data with anyone in Iran or North Korea. Blumenstock once found that Planet imagery he wanted for a project in Afghanistan was unavailable owing to an unspecified security concern. Identifying individual people or vehicles is impossible, Navulur says; this alleviates some privacy concerns, although pictures can be sharp enough to make out houses and other structures. (Of course, for large areas of the world, so is Google Maps' public imagery.)

KNOW YOUR NEEDS

Use of commercial images can also be restricted. Scientists are free to share or publish most government data or data they have collected themselves. But they are typically limited to publishing only the results of studies of commercial data, and at most a limited number of illustrative images.

Many researchers are moving towards a hybrid approach, combining public and commercial data, and running analyses locally or in the cloud, depending on need. Weiss still uses his tried-and-tested ArcGIS software from Esri for studies of small regions, and jumps to Earth Engine for global analyses.

The new offerings herald a shift from an era when scientists had to spend much of their time gathering and preparing data to one in which they're thinking about how to use them. "Data isn't an issue any more," says Roy. "The next generation is going to be about what kinds of questions are we going to be able to ask?" ■

Gabriel Popkin is a freelance writer in Mount Rainier, Maryland.

CORRECTION

The Careers feature 'Behind the scenes' (*Nature* **556**, 525–527; 2018) erred in its description of the winning photo. Callie Veelenturf was measuring pH, conductivity and temperature after the turtle had laid her eggs, not taking samples beforehand.

PUBLISHING

Unequal authorship

An analysis of more than 10 million scientific and medical studies published between 2002 and 2018 suggests that male authors will continue to outnumber female authors for at least the rest of this century. The findings, published in *PLoS Biology*, examined articles in science, technology, engineering, mathematics and medicine (L. Holman *et al.* *PLoS Biol.* **16**, e2004956; 2018). Female authors were particularly scarce in the fields of physics, computer science, maths and surgery. For example, women accounted for just 13% of prestigious last-author spots in physics studies. That percentage has crept upwards by about 0.1% a year since 2002, suggesting that if current trends hold, authorship in physics studies could reach equality in roughly 260 years. The gender disparity in authorship was especially pronounced in papers from Japan, Germany and Switzerland, whereas the most gender-equitable countries were in South America, Africa and elsewhere in Europe. "Without novel interventions, these fields are likely to remain gender-biased for many decades," says lead author Luke Holman, an evolutionary biologist at the University of Melbourne in Australia. "Despite recent gains, we still have far to go."

CONDUCT

Drop harassers

An online petition is calling for the US National Academy of Sciences (NAS) to revoke the membership of scientists who commit sexual harassment or assault. The move brings fresh attention to a troubling issue. As of 23 May, the petition, created by neuroscientist BethAnn McLaughlin of the Vanderbilt Kennedy Center in Nashville, Tennessee, had gained 2,289 signatures. "I'm impressed by the response," McLaughlin says. "It speaks to the fact that time's up in academia." Comments on the petition point to other non-scientific organizations that have revoked membership for misconduct. In February, the US National Science Foundation (NSF) announced that it would require universities and institutions to disclose the identities of NSF-funded researchers who have been disciplined for harassment. The NAS takes harassment and assault "very seriously," says Jennifer Walsh, a spokesperson for the National Academies of Sciences, Engineering, and Medicine. A 22 May statement by the presidents of each of the three academies says that a dialogue has started about the standards of professional conduct for membership.

MASQUES

Body of evidence.

BY MIKE ADAMSON

Murder has almost vanished from the spectrum of human existence. Not from any elevated moral sense or spiritual refinement, but merely because it is no longer true that dead men tell no tales.

The afternoon is bright and warm, and the terraces of the glittering vertical architecture of New Johannesburg are thronged with passers-by, robots going about their tasks. I stroll in the sun, enjoying the breeze from the sea, and the peace seems inviolate. I barely remember my original name — I must consult my files, but at the moment the detail is not important. There is someone I must meet, but my heart is ambivalent. Will I thank this person, or twist his head off with the hydraulic power hidden beneath svelte, ebony arms?

Towards the end of the twenty-third century, the human race has spread across the stars, our vessels have touched more than a hundred new worlds, and Old Earth is shaking free of the devastation of its past. The population is a fraction of its old levels, and technology has redressed many a shortcoming. Among the utopic inventions of a race new-born to the Universe are the many strategies by which death has ceased to be a certainty.

Death came for me in an accident — the failure of a sub-runner, one of the bullet trains that streak across the world by magnetic induction in vacuum tunnels. The passengers were saved by the many failsafes, but I was unlucky enough to be in the way of a collapsing stanchion as the mass of the passing Cairo–New Jo’burg express shook plascrete and reinforced alloys to pieces. I was told afterwards that there had been no time in which to suffer. A quick, clean exit.

It was only later that I discovered there had also been doubt. My body was damaged, but not beyond local systems to suspend and repair. Eyebrows were raised in some circles, reviews promised, but I heard nothing more; until a concerned doctor passed me the confidential finding that the supervising medical technician had not pursued the possible options but had selected immediate cyber-incarnation.

➔ **NATURE.COM**
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

Cyber-incarnation is the standard fallback by law, in which the personality is

downloaded from the chemical matrix of the brain and stored, pending reanimation in any of the available prostheses — cyborg, android, hardlight hologram, whatever the individual prefers. Only the rich can afford their own new body, of course; the rest of us make do with a time-share arrangement, and



although it's sometimes inconvenient, there are consolations.

Yesterday, I was a surfer riding high, glassy walls coming down on the ocean beaches; the day before, I delighted in music, accessing hardwired routines in the hologram I phased with — I had never played piano before but certainly want to again. Today, I am in a body I would have envied before my untimely passing, long and dark, and dressed to impress in flowing blues. She draws eyes and I revel in the fact that no one knows who I am. Only the dead recognize each other, sensing the simulacra we have become, and trade knowing smiles as we pass, for we are now an elite.

I see the one I'm after, taking a seat at an outdoor café at the foot of the cliff of glass fronting a hotel, its landing pads high above throwing cool shade. This is his lunch hour.

I walk with confidence, the strike of my heels almost lost in the sounds of strolling people, until I slide into a seat opposite and hold his eyes silently. He is not unhand-some, fair hair thick above a strong-boned skull with firm jaw; indeed his rejuvenation is superb, retarding his 120 years to a

comfortable 30, and I sense he appreciates this glove of flesh and alloys I wear. "Doctor Rensburg," I begin, a statement.

A robowaiter hovers nearby. Rensburg softly orders tea — for two.

"You have me at a disadvantage," he murmurs. "Isn't that how they used to put it?"

I offer a fine hand before I smile and at last recheck — Onika, yes, that was my name. "Onika Kabila."

He frowns, clearly not recalling. "Have we met?" He grins now. "I'm sure I would remember one so charming."

I flash him the smile his words earn, then sit back to stare off at the sea as the drone returns with fine china and a glass pot. When we cradle cups of fragrant blend, I have made up my mind. I will not kill him.

Instead I raise my cup in salute. "Thank you," I whisper, my smile suddenly very genuine. "You could have dropped me back into my living body but you prematurely reassigned me. I know you earn a gratuity from the cybernetics manufacturers for every patient you send into hard-backup, keeping the android industry in full swing, and for a long time I meant to kill you for

it." His eyes widen and my electronic senses see pulses race, pupils dilate, a dozen other tells. "But they would only pour you back onto another hard drive, and I'm not so hate-filled I would destroy your brain entirely to prevent it." My stare holds his eyes like glass knives as he pivots upon the uncertainty of the instant. "But you know what? I like where I am now. It's not so different — and better in some ways." I sip again, then rise, lean across the table and speak softly by his ear. "So I'll wish you good luck and long life, but hope you'll leave the choice to your patients in the future."

I rise and walk on in the sun, losing myself among the crowd on the terraces, and know Rensburg is trembling with shock, a very human thing of which I am no longer capable. But that's a small price to pay.

Because I have learnt that life really does begin at the end. ■

Mike Adamson holds a PhD in archaeology from Flinders University of South Australia. After early aspirations in art and writing, Mike returned to study and currently lectures in anthropology.

ILLUSTRATION BY JACEY

natureOUTLOOK

HUNTINGTON'S DISEASE

31 May 2018 / Vol 557 / Issue No 7707



Cover art: Neil Webb

Editorial

Herb Brody,
Michelle Grayson,
Richard Hodson,
Elizabeth Batty

Art & Design

Mohamed Ashour,
Andrea Duffy,
Wesley Fernandes,
Wojtek Urbanek

Production

Ian Pope, Karl Smart

Sponsorship

David Bagshaw,
Judy Yeh

Marketing

Shan Li,
Nicole Jackson

Project Manager

Rebecca Jones

Art Director

Kelly Buckheit Krause

Publisher

Richard Hughes

Editorial Director

Stephen Pincock

Magazine Editor

Helen Pearson

Editor-in-Chief

Philip Campbell

There are many things that parents would love to pass down to their children. Houses, jewellery, money — all can be welcome gifts from one generation to the next. But not everything that can be bequeathed is so desirable. Huntington's disease — a neurodegenerative condition that causes uncontrollable movements, emotional disturbance and the loss of mental abilities — is an especially unfortunate genetic hand-me-down.

At a glance, the biology of Huntington's disease seems to be simple (see page S36). The condition has been traced to a mutation in a single gene on chromosome 4 that is responsible for producing a protein called huntingtin. And yet fundamental aspects of the molecular and cellular processes that underlie Huntington's disease remain a mystery. Glaringly, researchers have still not worked out huntingtin's role in the cell.

There is no cure on the horizon. But a clinical trial of a potential treatment has raised hopes that such research is starting to bear fruit. The innovative treatment comprises an antisense oligonucleotide — a molecule that binds to messenger RNA to prevent the production of a specific protein (S39). Another possibility is to edit the faulty gene directly (S42). And as researchers pursue these new approaches, they must also wrestle with how to measure success without the need to track the health of patients for years, or even decades, to come (S46).

Huntington's disease is unusual in that its diagnosis generally occurs well into the child-bearing years, which can lead to anguished decisions over whether to take a high-stakes genetic gamble with future offspring (S38). But in rare instances, the disease also strikes children (S44).

We are pleased to acknowledge the financial support of F. Hoffmann-La Roche in producing this Outlook. As always, *Nature* has sole responsibility for all editorial content.

Herb Brody

Chief supplements editor

CONTENTS

S36 BIOLOGY

Chain of mystery

Pinning down the molecular cause of Huntington's disease

S38 GENETIC TESTING

Darkness and light

One man's search for certainty

S39 TREATMENTS

The big hope for Huntington's

Antisense oligonucleotides hit clinical trials

S42 GENE EDITING

To cut is to cure

Could CRISPR hold the key to treating Huntington's disease?

S44 PAEDIATRICS

Ahead of time

Huntington's disease in the young

S46 CLINICAL TRIALS

The endpoint is near

Ways to improve comparisons of potential treatments

S48 RESEARCH

4 big questions

Balancing the need for basic research with innovation

Nature Outlooks are sponsored supplements that aim to stimulate interest and debate around a subject of interest to the sponsor, while satisfying the editorial values of *Nature* and our readers' expectations. The boundaries of sponsor involvement are clearly delineated in the *Nature Outlook* Editorial guidelines available at go.nature.com/e4dwzw

CITING THE OUTLOOK

Cite as a supplement to *Nature*, for example, *Nature* Vol. XXX, No. XXXX Suppl., Sxx–Sxx (2018).

VISIT THE OUTLOOK ONLINE

The *Nature Outlook Huntington's disease* supplement can be found at www.nature.com/collections/huntingtons-disease-outlook. It features all newly commissioned content as well as a selection of relevant previously published material that is made freely

available for 6 months.

SUBSCRIPTIONS AND CUSTOMER SERVICES

Site licences (www.nature.com/libraries/site_licences): Americas, institutions@natureny.com; Asia-Pacific, <http://nature.asia/jp-contact>; Australia/New Zealand, nature@macmillan.com.au; Europe/ROW, institutions@nature.com; India, npgindia@nature.com. Personal subscriptions: UK/Europe/ROW, subscriptions@nature.com; USA/Canada/Latin America, subscriptions@us.nature.com; Japan, <http://nature.asia/jp-contact>; China, <http://nature.asia/china-subscribe>; Korea, www.natureasia.com/ko-kr/subscribe.

CUSTOMER SERVICES

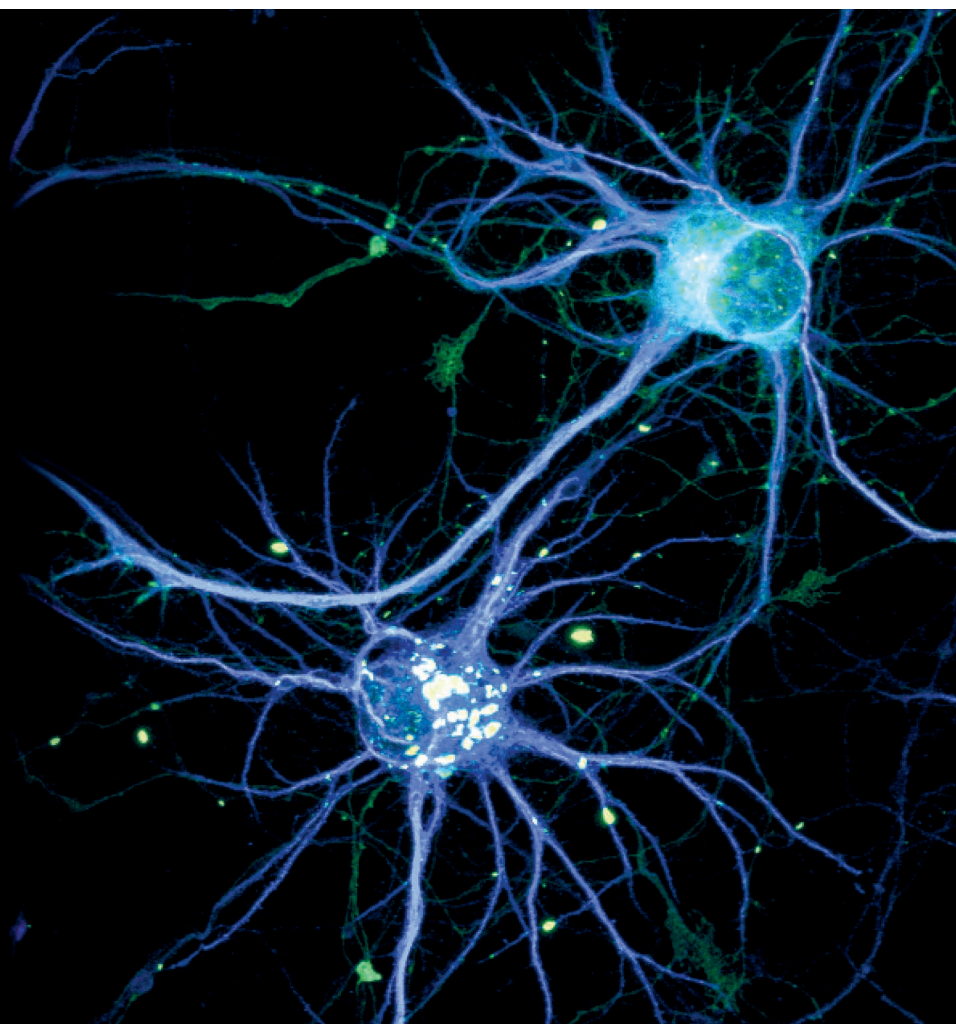
Feedback@nature.com

Copyright © 2018 Macmillan Publishers Ltd. All rights reserved.

BIOLOGY

Chain of mystery

Decades after uncovering the genetic basis of Huntington's disease, researchers remain puzzled by the condition's molecular cause.



HTT exon 1 (green), a shortened form of the protein huntingtin, which is implicated in Huntington's disease, forms clumps in embryonic rat neurons.

BY SARAH DEWEERDT

In many respects, Huntington's disease is the epitome of a well-understood hereditary disease. The condition was comprehensively described almost a century and a half ago¹. Its symptoms are well-characterized: involuntary, jerky movements known as chorea; difficulty in coordinating voluntary movements; cognitive impairment; and psychiatric issues such as changes in mood. And its pattern of inheritance is clear — a person with a parent who is affected has a 50% chance of developing the disease.

Researchers know exactly where to find the gene that is implicated in Huntington's disease. Known as *HTT*, and located near the tip of the short arm of chromosome 4, it was the first gene to be pinned to a chromosome region through genetic mapping techniques, in 1983 (ref. 2). Ten years later, at the dawn of the genomic era, scientists homed in on the sequence of *HTT*³. The mutation that causes the disease is now well-understood: an abnormal expansion of a repetitive sequence of DNA comprising a triplet of bases — cytosine (C), adenine (A) and guanine (G) — towards one end of the gene. A person who carries a copy of *HTT* containing

40 or more of these triplets, or CAG repeats, will develop the characteristic symptoms of Huntington's disease, typically around the age of 45, and then succumb to the condition within about two decades of the onset of motor problems. A person who has between 36 and 39 such copies could develop Huntington's disease, but might not. Someone with 35 or fewer copies of the CAG repeat will not develop the disease.

And yet, fundamental aspects of the biology that underlies Huntington's disease remain a mystery. "Many genetic diseases are relatively straightforward at the molecular level," says Ronald Wetzel, a structural biologist at the University of Pittsburgh in Pennsylvania. "The frustration we have with understanding Huntington's disease comes, in part, because of this background of our more-positive experience with other disease mechanisms."

For starters, the function of huntingtin, the protein encoded by *HTT*, isn't fully understood. More specifically, researchers don't know what the portion affected by the mutation does. And they aren't sure why the mutant protein causes problems in cells, how those problems begin, or which of several forms of the protein is responsible.

As a promising potential treatment for Huntington's disease — designed to silence the expression of *HTT* with a synthetic molecule known as an antisense oligonucleotide that binds to messenger RNA — moves into clinical trials (see page S39), answering basic questions such as these becomes more important than ever. With such progress being made, ignorance of the disease's biological basis should not be allowed to create a bottleneck for research. Knowing more about how the mutant protein leads to damage in nerve cells, in particular, could be crucial for the success of treatments for the condition. "You need to know how to most-selectively target that mRNA," says Matthew Disney, a biochemist at the Scripps Research Institute in Jupiter, Florida.

TRIPLET TROUBLE

The DNA sequence CAG encodes the amino acid glutamine. The CAG repeats in *HTT* therefore lead to the production of a string of glutamines, known as a polyglutamine chain, which is abnormally long in people with the large numbers of repeats that are associated with Huntington's disease. But the function of the chain, and the reason for the existence of

KENNETH W. DROMBOSKY, RONALD WETZEL & TUA C. JACOB

the CAG repeats, is unknown.

A somewhat provocative explanation, proposed by Chiara Zuccatto and colleagues at the University of Milan in Italy, is that Huntington's disease is the unfortunate by-product of an evolutionary advantage. The number of CAG repeats in *HTT* varies among species of vertebrate, and their expansion is greatest in humans. Huntingtin is essential for the development of the nervous system before birth; indeed, the researchers contend that CAG repeats might have contributed to the development of the complexity of the vertebrate brain⁴.

But other researchers say that CAG repeats are just as likely to have arisen by chance, and that they have little effect until a threshold is crossed. "These sequences are difficult to copy," says Cynthia McMurray, a biochemist at Lawrence Berkeley National Laboratory in California, "so sometimes the polymerase [the enzyme responsible for copying DNA] stutters around, and it can't get through it." This results in extra copies being incorporated into the DNA, lengthening the repeat.

Ten hereditary diseases are known to be caused by the expansion of CAG repeats in various genes, and other triplet-repeat diseases exist — CAG repeats just happen to be particularly vulnerable to these polymerase 'stutters'.

It's not just the purpose of the polyglutamine chain that is difficult to unpick. The function of huntingtin is poorly understood and the protein might have multiple roles in cells. Although it is expressed most strongly in the brain, huntingtin is found throughout the body and has been shown to interact with more than 100 proteins. "That's sort of amazing," says Stefan Kochanek, a gene-therapy specialist at University Hospital Ulm in Germany. "It's a very large number of proteins."

Some experiments suggest that huntingtin helps to transport proteins around the cell. Other findings hint that it might play a part in transcription. It might also aid the proper folding of proteins, or help proteins to form complexes with one another. "The idea of there being one prime mechanism for what huntingtin protein does is far from clear," McMurray says. "It does a lot of things."

Scientists have documented numerous biochemical abnormalities in animal models of Huntington's disease, as well as in people with the condition. "There are probably more things you can measure that are going off the rails in Huntington's than stay on the rails," says David Housman, a biologist at the Massachusetts Institute of Technology in Cambridge, who was involved in identifying *HTT*. The balance between protein synthesis and degradation, the function of cellular structures such as the endoplasmic reticulum, which is involved in protein processing, and the cell's responses to stress, for example, become disrupted. But Housman says that the cause-and-effect relationship between these observations, or what goes awry first, is unclear. That makes it difficult to glean clues about huntingtin's normal

function from observations of what goes wrong in Huntington's disease — as well as to design a treatment to target the initial steps in the development and progression of the condition.

An analysis of the structure of huntingtin⁴ published in March supports the idea that it is some kind of molecular chaperone — a protein that aids the function of other proteins — says Kochanek, who led the work. It took a decade to come to fruition, he adds, because the protein's extreme flexibility made it difficult for researchers to get clear images of the structure.

The analysis does not, however, clearly reveal the structure of the polyglutamine chain, so it cannot provide further information about the function of that part of the protein. Some researchers suspect that the portion of huntingtin in which the disease-causing mutation occurs might have little involvement in the protein's normal function — a molecular curlicue, just along for the ride.

But others contend that it must do something, even though it's not clear what. Polyglutamine chains often facilitate interactions between proteins, and mice that lack this portion of *HTT* show behavioural and biochemical changes. So the question remains: how exactly does mutant huntingtin cause such problems?

"The idea of there being one prime mechanism for what huntingtin protein does is far from clear."

STICKY CHAINS AND CLUMPS

Scientists have debated whether the damage to cells in Huntington's disease stems from a loss of the normal function of huntingtin, a toxic effect that is unique to the mutant protein, or a combination of both. There's also disagreement on which form of the mutant protein is the culprit — a question that is difficult to answer because many versions of huntingtin exist in the cell at the same time, owing to splicing and modification processes. "It's like having too many viable suspects in an Agatha Christie novel," Wetzel says.

Some researchers think that full-length huntingtin is the prime suspect. Chains of glutamine tend to be sticky, McMurray says, and huntingtin containing too many glutamines in a row might adhere more strongly to other molecules than it would normally — gumming up the movement of proteins through the cell. She suggests that the resulting impairment of huntingtin's transportation activity would disrupt metabolism and other cellular functions gradually, which is consistent with the slow progression of Huntington's disease.

But many others instead point to a shortened form of huntingtin, known as *HTT* exon 1, in which exon 1 refers to the protein-coding region of *HTT* that contains the CAG repeats. It's found only in people with Huntington's disease and contains sequences that are not usually translated into protein. *HTT* exon 1 could be more toxic than any other form of

huntingtin. For example, mouse models in which only *HTT* exon 1 is expressed show all the main features of Huntington's disease⁵. Meanwhile, studies in fruit flies suggest that of all huntingtin fragments produced by the cell, *HTT* exon 1 is the most harmful⁶.

If *HTT* exon 1 is the real cause of Huntington's disease, exactly how it exerts its damaging effect remains unclear. That's because, compared with flexible full-length huntingtin, *HTT* exon 1 is a downright shape-shifter. It switches freely and rapidly between multiple conformations, making it almost impossible to isolate the effects of any particular one.

Researchers have several hypotheses. One idea is that, similar to how full-length huntingtin containing many glutamines might gum up cells, the longer the polyglutamine chain of *HTT* exon 1 is, the more strongly the fragment can bind to other molecules. Other scientists blame the tendency of *HTT* exon 1 to aggregate with itself. Nerve cells from people with Huntington's disease contain clumps of huntingtin-related proteins, as well as smaller fibrils comprising thousands of copies of *HTT* exon 1. In turn, such aggregations can disrupt a variety of cellular functions, leading inexorably to neurodegeneration.

Similar clumps are found in other neurodegenerative conditions such as Alzheimer's disease, Parkinson's disease and amyotrophic lateral sclerosis. They also occur in all known diseases caused by expanded CAG repeats, including spinocerebellar ataxia.

Wetzel and his collaborators have shown that polyglutamine chains acquire a much greater propensity to stick to each other and form aggregates when they reach 37 amino acids long. This observation provides a potential link between the disease-causing threshold of the CAG repeats and the biochemistry of huntingtin. Still, others caution that huntingtin clumps might be the result, not the trigger, of the processes responsible for the slow progression of Huntington's disease.

It's tempting to think that such debates won't matter if treatments such as the antisense oligonucleotide in clinical testing can simply turn off huntingtin expression altogether. But some of the antisense molecules being tested target a portion of *HTT* that is distant from exon 1, Disney says. "Those antisense oligonucleotides are going to ablate full-length huntingtin, but they are not going to affect this mini version of huntingtin," he says. Perhaps it's a hint that huntingtin might be too complex for such simple solutions. ■

Sarah DeWeerd is a freelance science writer in Seattle, Washington.

1. Huntington, G. *Med. Surg. Rep.* **26**, 317–321 (1872).
2. Gusella, J. F. *et al. Nature* **306**, 234–238 (1983).
3. The Huntington's Disease Collaborative Research Group. *Cell* **72**, 971–983 (1993).
4. Guo, Q. *et al. Nature* **555**, 117–120 (2018).
5. Mangiarini, L. *et al. Cell* **87**, 493–506 (1996).
6. Barbaro, B. A. *et al. Hum. Mol. Genet.* **24**, 913–925 (2015).



Mark Newnham chose to discover whether he carries the gene mutation for Huntington's disease.

GENETIC TESTING

Darkness and light

The confirmation that a person will develop Huntington's disease can bring them more uncertainty — but also relief.

BY SIMON ROACH

Mark Newnham has seen the future, and it's etched on his father's face. Despite being in good health, the 31-year-old knows that Huntington's disease is coming — he just doesn't know when.

Like many people living under the shadow of the condition, Newnham, who lives in London, first heard about Huntington's disease when it struck older members of his family. A great-uncle had been diagnosed with it at the end of his life. So when Newnham's father started to develop the involuntary movements associated with the condition, he got tested for the gene mutation responsible. His father's diagnosis meant that Newnham — who was 20 years old at the time — had a 50% chance of carrying the gene. "I didn't know what Huntington's disease was when my Dad told me that he had it," Newnham says.

In the ten years since, his father's symptoms have progressed to include more severe involuntary movements, memory difficulties and mood swings. Later, his driving became worryingly erratic. Thinking about those years, in which his father's mental health began to decline, is painful, Newnham says — he feels as though he has been witnessing someone "at war with himself, every day".

Throughout his early twenties, and despite his father's illness showing him what might await, Newnham did not want to take the genetic test that would reveal whether he

had inherited the mutation for Huntington's disease. "I was more of a free spirit," he says. "I thought, 'I don't need to know.' I can get on with it and just see if it happens later on in life."

That attitude reflected his general approach to life. As an actor and musician, he launched from one project to the next with little thought about what would come later. "I wanted to headline the Glastonbury Festival, and I wanted to become the next Johnny Depp," he laughs. "Those were my goals."

He continued on that path, he says, until he met his partner. Finding happiness and stability changed his perspective on Huntington's disease — especially when the couple thought about having children. Could he face rolling the dice when he might pass the condition to his offspring?

Newnham sought genetic counselling through the UK National Health Service, during which he explored the impact that testing could have on his life. This involved considering his motivation for being tested, as well as changes that he might need to make in the event that he did have the mutation. Aside from the emotional strain that such testing can bring, it also raises questions about physical care and finances; the certainty of knowing you have the mutation can make it more difficult to get long-term health or life insurance. Genetic counsellors can help those who might be affected to pick through the entangled pros and cons.

After three sessions, and given his and his partner's desire to have children, Newnham

concluded that he needed to know his status with respect to Huntington's disease. "We didn't want to have a child without that certainty," he says. The result was not what he had hoped for. Like his father, he carries the mutated gene.

Newnham is yet to experience symptoms, and it could be decades before he shows signs of the disease. He still works as an actor and musician, but says that his priorities have changed. "The test results made me realize that what really drove me as a person before, and what ambitions I had, they're not as important now," he says. The dream of performing at Glastonbury will never be gone, but spending time with family and friends seems more important. This shift in perspective has given him a quiet contentment, he adds.

More willing to look ahead, Newnham and his partner immediately began to explore how they could have a child who would not carry the Huntington's disease mutation.

"I wanted to make sure that I don't pass this on to the next generation," he says. That meant going through a process called preimplantation genetic diagnosis (PGD).

In PGD, embryos created through *in vitro* fertilization (IVF) are screened for specific genetic disorders; only those without the related mutations are implanted. In England, up to three rounds of PGD are available at no cost to people who meet certain criteria. (In the United States, many health-insurance plans won't cover the process, so people typically pay US\$15,000–25,000 for IVF with PGD.)

Not everyone with a family history of Huntington's disease goes to these lengths. Some leave it to chance. And various religious groups have reservations about prenatal genetic-screening methods such as PGD because any embryos found to have genetic abnormalities will be destroyed.

For those who do elect for PGD, the chances of success are low — as Newnham and his partner found out when they received the news that their journey towards parenthood had, for now, come to an end. The IVF part of the process, itself a complicated procedure, had failed and there were no embryos to test.

For now, the couple are weighing up the options. Adoption is a possibility, but people who will go on to develop conditions such as Huntington's disease tend to be at the bottom of the list because of their own care needs later in life, Newnham says.

Despite the prospect of a life without children, Newnham does not regret his decision to get tested. At least, he explains, he is moving forward with his eyes open. And advances in research fill him with "immense hope" that some form of treatment will be available in his lifetime — too late for his father, perhaps, but soon enough that the risk of having a child with Huntington's disease might no longer be one of life or death. ■

Simon Roach is a freelance writer in Glasgow, UK.

TREATMENTS

The big hope for Huntington's

A quarter of a century after its discovery, researchers are finally unlocking ways to neutralize the gene behind Huntington's disease.

BY LIAM DREW

In September 2014, at a meeting of the European Huntington's Disease Network, Sarah Tabrizi announced the launch of a drug trial. Tabrizi, a neurologist and director of the University College London Huntington's Disease Centre, would be working with Ionis Pharmaceuticals of Carlsbad, California, to test the safety and tolerability of a drug candidate called IONIS-HTT_{Rx} in people for the first time. The drug had been designed to reduce the amount of protein being made by the gene that causes Huntington's disease.

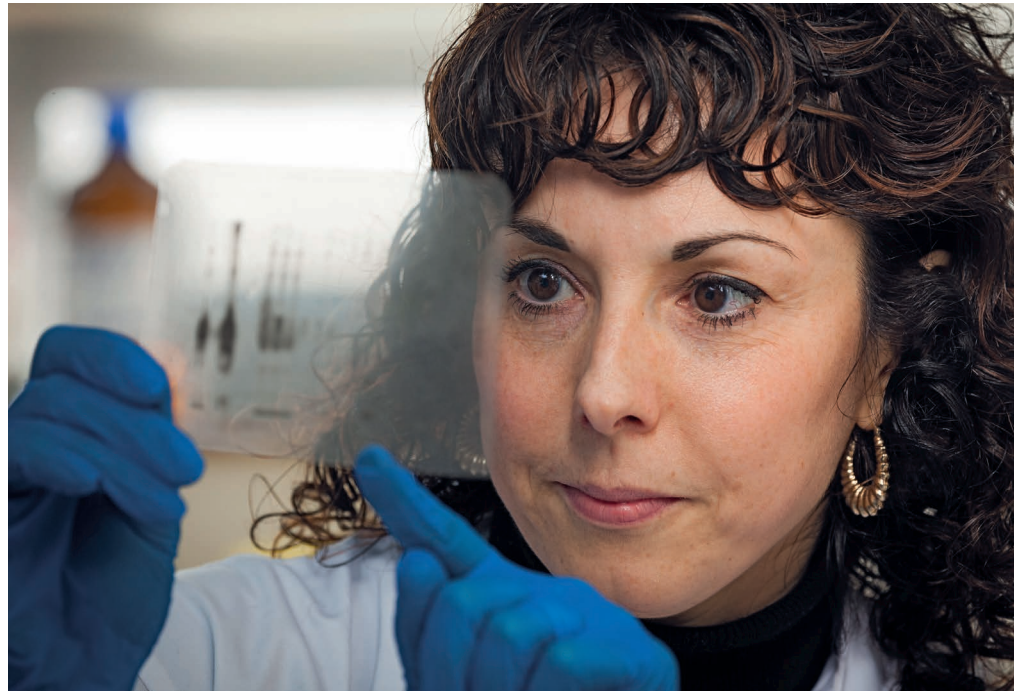
That gene is huntingtin (*HTT*). Inheriting just one mutated copy brings about a progressive neurodegeneration that typically begins in a person's forties. The condition's most distinctive symptom is involuntary, jerky limb movements. This is preceded by subtler psychiatric symptoms and followed by a disabling dementia.

IONIS-HTT_{Rx} is an antisense oligonucleotide (ASO): an artificial chain of 12–25 nucleotides that is designed to prevent the production of protein from a specific gene. In the trial, people in the initial stages of Huntington's disease would receive four monthly injections of ASO directly into their cerebrospinal fluid (CSF) through a lumbar puncture. IONIS-HTT_{Rx} was expected to diffuse into the brain, where it would suppress the production of the protein huntingtin in neurons. As well as ensuring that the drug had no adverse effects, Tabrizi and Ionis would use a new assay to measure levels of mutant huntingtin.

The Huntington's disease community was excited about the trial — a way of silencing *HTT* has been sought since the gene was discovered in 1993. But the path to using ASOs as treatments had been rocky and the brain is notoriously difficult to target with drugs. Tabrizi recalls that after her presentation, colleagues told her, "It'll never reach the brain. It's never going to work."

In December 2017, however, a press release revealed that the doubters had been wrong — the trial had been a success. And in March 2018, Tabrizi unveiled the resulting data at the final session of the 13th Annual Huntington's Disease Therapeutics Conference.

Her most important slide showed decreases in the level of mutant huntingtin in trial participants' CSF — indicative of reduced levels of the toxic protein in their brains — that were proportional to the amounts of drug the volunteers had received. At the two highest doses,



Sarah Tabrizi is helping to move innovative potential treatments for Huntington's disease into the clinic.

production of the protein had, on average, decreased by about 40%.

"People started crying," says Jeff Carroll, a neuroscientist who investigates Huntington's disease at Western Washington University in Bellingham. "Everybody who works in Huntington's disease long enough meets families and

"It'll never reach the brain. It's never going to work."

gets to know them, so it becomes very personal." Carroll's connection to his work runs particularly deep. He began his career in neuroscience after his mother was diagnosed with Huntington's disease. Then, in 2003, he discovered that he, too, had the mutation for the condition. Looking at Tabrizi's slide, Carroll thought, "This is a graph that is changing my life."

Tabrizi emphasizes that the trial did not show that IONIS-HTT_{Rx} is able to treat Huntington's disease. It demonstrated only that the drug was safe and well-tolerated, and — crucially — that it engaged its target in the brain. "What we now have to do," she says, "is to move quickly forward to larger, longer trials to test whether the drug slows disease progression."

These trials will be run by pharmaceutical company Roche of Basel, Switzerland.

In 2013, it partnered with Ionis to develop IONIS-HTT_{Rx} and, after the initial trial, Roche acquired the drug for US\$45 million. The companies will continue to collaborate. If all goes to plan, a large phase III trial of IONIS-HTT_{Rx} will commence later in 2018.

Assessing where the project stands at present, Tabrizi says she has become fond of a quote from a speech by Winston Churchill: "Now, this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning."

THE BEGINNING OF THE BEGINNING

The molecular basis of ASO technology is the stuff of secondary-school textbooks. In double-stranded nucleic acids, the base guanine binds to cytosine, and thymine (in DNA) or uracil (in RNA) binds to adenine. This pairing enables DNA both to replicate and to supply cells with instructions for making protein.

ASOs are designed to be complementary to the messenger RNAs of specific genes, which act as templates for protein production. When a cell is flooded with a particular ASO, the ASO will bind to its target mRNA, preventing it from guiding protein synthesis.

Forty years ago, researchers at Harvard University in Cambridge, Massachusetts, demonstrated¹ that an ASO made of DNA

could stop the replication of a virus by blocking viral protein production. And a year later, in 1979, it was shown² that the binding of ASOs to mRNA not only prevented protein synthesis, but also triggered the degradation of mRNA.

Given that ASOs could, in principle, suppress the expression of any gene, these findings raised hopes of a fresh approach to the treatment of many diseases. And conditions such as Huntington's, caused by faulty genes that produce proteins with toxic effects, were seen as particularly promising targets.

There was, however, a considerable problem: DNA makes a lousy drug. A good drug tends to distribute evenly through the body, so that sufficient amounts reach the desired target. To achieve this, the drug must persist for a substantial amount of time. On this score, ASOs

"People were so enthralled by genetics. They thought if you had the gene, you had the cure."

face a major problem when given to mammals, which produce high levels of enzymes called nucleases that digest nucleic acids.

Also, to be effective, ASOs must bind to target mRNAs both tightly and with specificity. Complementary base pairing means that ASOs have preferred-partner mRNAs. However, because they are highly charged molecules, ASOs can also bind to mRNAs to which they are not perfectly complementary — thereby affecting the expression of other genes — as well as to proteins. Both events could give rise to off-target effects and toxicity issues.

Making ASO-based treatments a reality has required the creation of molecules that retained nucleic acid's property of complementary base pairing, while otherwise overhauling ASO chemistry.

Ionis has been a central player in this pursuit since its foundation in 1989. Frank Bennett, vice-president of research, set up the company's laboratories and later led the development of IONIS-HTT_{Rx}. He stresses that Ionis was "founded on an idea for a technology". Unlike most start-ups, it did not license technology from elsewhere. And the technology that it, like other companies developing ASOs, has produced has undergone a substantial evolution.

Initially, the company focused on treatments for viral infections and cancer. It used first-generation ASOs, the backbones of which had already been chemically modified to differ from those of DNA, which fractionally increased their resistance to nucleases. In 1998, Ionis's drug fomivirsen (Vitravene) became the first ASO to be approved by the US Food and Drug Administration. It was injected into the eyes of people with AIDS to treat a viral infection that can cause blindness in those with compromised immune systems.

Jeffrey Carroll's research is motivated by a personal connection to Huntington's disease.

But after advances in treating HIV infection that maintained the immune response, the drug fell out of use. Other early ASOs were clinical failures. In the mid-1990s, several companies who had invested heavily in ASO technology withdrew from the field.

The frustration extended to basic research. The first published report of an attempt to suppress the production of huntingtin in mice using ASOs described a failed experiment³. That study's lead researcher, Ole Isacson, who works on neurodegeneration at the Harvard Stem Cell Institute in Cambridge, recalls the conflicted time that followed the description of *HTT* in 1993. "People were so enthralled by genetics," he says. "They thought if you had the gene, you had the cure. Those of us with direct experience of working on disease models didn't get that feeling."

Bennett concurs. The discovery of *HTT* piqued his interest in Huntington's disease, but "at the time, we weren't ready," he says. Instead, Ionis focused on improving the stability of ASOs. Second-generation ASOs were developed in the late 1990s and early 2000s by incorporating further chemical modifications that increased resistance to digestion by nucleases. Isacson says that the enhanced stability of present ASOs, which can act for months, compared with the short-lived molecules that he used in 1997, is "one of the most remarkable improvements in technology that I've seen". Only after Ionis had developed such ASOs did Bennett begin to tackle genetic disorders affecting the brain.

To do so, Ionis forged a collaboration, in 2003, with Don Cleveland, a neuroscientist at the University of California, San Diego, that aimed to treat a genetic form of amyotrophic lateral sclerosis, or motor-neuron disease. Then, in 2006, buoyed by progress they had made using mouse models of that condition, Ionis and Cleveland began to work on Huntington's disease. In 2012, they published studies⁴ showing

ASOs that target *HTT* mRNA could reverse Huntington's-disease-like symptoms in mouse models of the condition, alongside a demonstration that ASOs downregulate huntingtin production in the brains of rhesus macaques.

After this proof-of-concept work, Ionis designed and validated an ASO that would work in people, established the best way to deliver it to the brain — optimizing the lumbar-puncture procedure, for example — and then determined the parts of the brain that the drug was most likely to reach. Finally, with Tabrizi, Roche and the CHDI Foundation (a US non-profit organization that funds research on Huntington's disease), it developed the assay⁵ to track levels of mutant huntingtin.

THE END OF THE BEGINNING

The optimism that surrounds IONIS-HTT_{Rx} stems from the well-established link between reduced levels of mutant huntingtin and improvements in symptoms in animal models of Huntington's disease. Yet some researchers are concerned that IONIS-HTT_{Rx} suppresses not only the production of mutant huntingtin, but also synthesis of the normal protein.

This is because selectively targeting mRNA from the mutated copy of *HTT* — leaving mRNA from the normal copy untouched — poses a huge technical challenge. The mutation that causes Huntington's disease is an overlong run of the nucleotide triplet CAG (see page S36): normal *HTT* contains 17–35 consecutive such triplets, whereas in people with the condition, at least one copy of the gene has 36 or more in a row. Consequently, an ASO of around 20 nucleotides that targets CAG repeats would bind to both normal and disease-causing versions of *HTT* mRNA. And, problematically, more than 50 other human genes also contain 10 or more CAG repeats, which means that targeting the sequence could produce unwanted side effects.

An alternative approach to targeting *HTT*



mRNA is being developed by Wave Life Sciences in Cambridge, Massachusetts. Its strategy takes advantage of functionally insignificant differences that can often be found between a person's two copies of *HTT*. If their disease-causing gene differs from their normal copy by a single nucleotide — an A, for example, instead of a C — an ASO can be designed to target only the *HTT* mRNA containing the substituted nucleotide, a mutation known as a single nucleotide polymorphism (SNP). “We could make those SNPs into therapeutic targets for drugs,” says Paul Bolno, Wave's chief executive.

Wave was founded on an innovative means of manufacturing ASOs, in which the intrinsic symmetry — or ‘handedness’ — of each nucleotide is specified during ASO synthesis. But the company's SNP-targeting approach to Huntington's disease also requires technical innovation in genotyping. If a physician were to prescribe a drug that suppresses a gene on the basis of it containing a particular SNP, he or she would need to be certain that the SNP is in the patient's disease-causing version: inadvertently suppressing the normal copy while leaving mutated *HTT* unaffected could accelerate the disease. In conventional gene sequencing, DNA from a person's two copies of a gene is combined — it's possible to discover which mutations the person has, but not on which chromosome (of the pair) they are found. To determine that in Huntington's disease, the sequencing reaction must follow the same strand of DNA from the region of CAG repeats to the SNP that is being used to differentiate between versions of the gene. Wave says that its sequencing platform does exactly this. But to meet regulatory approval, the error rate will have to be essentially zero.

The company has now begun separate phase I trials of two ASOs that each target one of the two most common SNPs in mutated *HTT*. The approach represents a personalized route to treating Huntington's disease — only people with the targeted SNPs can benefit. Unfortunately, about 30% of those with the condition have neither SNP. Bolno says that Wave is looking for further SNP targets.

The need for specificity is contentious. Bolno points to studies in mice, in which switching off the production of normal huntingtin causes deleterious effects, as evidence that suppressing both forms of huntingtin in people might have unwanted consequences.

Ionis and Tabrizi, however, disagree. Although studies in mice show that normal huntingtin is crucial for early development, they emphasize that, in adult animals, its function is much less important. Besides, they point out, ASOs do not reduce levels of huntingtin to zero. They cite other studies in mice in which lowering but not totally removing huntingtin had no adverse effects. What's more, they say, rhesus macaques given IONIS-HTTRx for up to nine months showed no detrimental effects.

Both Ionis and Wave are working with Carroll to resolve this pivotal issue. Carroll

RNA INTERFERENCE

Another way to halt huntingtin

Antisense oligonucleotides (ASOs) are not the only way to suppress protein synthesis by targeting messenger RNAs. Cells already have an intrinsic mechanism for stifling mRNA that involves producing short molecules of RNA that are complementary to mRNAs. When a short RNA binds to an mRNA, protein synthesis is prevented through a phenomenon known as RNA interference (RNAi).

Most approaches to harnessing RNAi involve introducing a gene into cells that then constantly produces a desired short RNA to permanently suppress synthesis of a target protein. Three companies — Spark Therapeutics in Philadelphia, Pennsylvania; UniQure in Lexington, Massachusetts; and Voyager Therapeutics in Cambridge, Massachusetts — are pursuing RNAi approaches to treating Huntington's disease.

Each is developing a virus-based vector — which is incapable of replicating and therefore poses no threat to health — that delivers a gene encoding a short RNA that inhibits huntingtin production. However, this one-off genetic treatment raises potential safety concerns. Permanently

modifying tissues to produce RNA might reduce the need for repeated dosing, but the procedure cannot be reversed if side effects develop. And, unlike ASOs, which need only to be introduced to the cerebrospinal fluid, these vectors must be administered in the vicinity of the cells to be treated. In Huntington's disease, researchers have focused on the striatum — an area in the middle of the brain that is most-visibly affected by the condition.

Controversy exists as to whether suppressing huntingtin in the striatum alone will halt Huntington's disease. This brain structure has an important role in the condition's progression, and Pedro Gonzalez-Alegre, a neurologist at the University of Pennsylvania's Perelman School of Medicine in Philadelphia, thinks that “improving one key brain region will have benefits beyond that area itself”. But ultimately, he concedes, “we will answer this question only when we do it in humans.” Perhaps, he suggests, the strong RNA-based suppression of huntingtin in the striatum might be supplemented with ASOs to more modestly suppress the protein throughout the brain and body. **L.D.**

despairingly cites two almost identical studies^{6,7} in mice that gave radically different results. “Right now, we don't know enough,” he says. “And you can only get so much from mice.”

Another question about the long-term future of ASOs concerns the plausibility of giving lumbar punctures to patients on a regular basis, potentially, for decades. ASOs that can cross the blood–brain barrier, which do not need to be introduced directly into the CSF, are still in early development. And there are methods other than a lumbar puncture for getting drugs into the CSF: some people with multiple sclerosis, for example, use implanted pumps for the task. Tabrizi says that finding an alternative delivery system is a “post-approval problem” — given that no current treatment stops the condition, she notes, people with Huntington's disease accept that they'll need to visit a hospital regularly to receive treatment.

Other potential genetic treatments hold the advantage of having to be administered only once. At the moment, the most viable such alternative is RNA interference (see ‘Another way to halt huntingtin’). Looking further ahead, the gene-editing tool CRISPR could correct *HTT* directly (S42). But although the enthusiasm that surrounds this technique is valid, the history of ASOs shows that much work is needed to move exciting ideas into the clinic.

“ASOs are reaching prime time,” says Tabrizi,

for Huntington's disease and, potentially, brain diseases in general. She can now smile about the doubters that she encountered on announcing the trial. “This is science,” she says. “When you're trying to develop new therapies, you're always going to have sceptics, and you have to just carry on with what you believe in. And I believed in this.”

Carroll, like many other researchers and people who have been touched by Huntington's disease, has been riding a fresh wave of hope since Tabrizi revealed results of the ASO trial in March. “Ever since I got my diagnosis,” he says, “I have operated under the assumption that I'll die at the same time my mum did — that I'll get sick when she did. Seeing that graph was the first time I've believed that it could be better.” ■

Liam Drew is a science writer based in London.

1. Zamecnik, P. C. & Stephenson, M. L. *Proc. Natl Acad. Sci. USA* **75**, 280–284 (1978).
2. Donis-Keller, H. *Nucleic Acids Res.* **7**, 179–192 (1979).
3. Haque, N. & Isacson, O. *Exp Neurol.* **144**, 139–46 (1997).
4. Kordasiewicz, H. B. *et al. Neuron* **74**, 1031–1044 (2012).
5. Wild, E. J. *et al. J. Clin. Invest.* **125**, 1979–1986 (2015).
6. Dietrich, P., Johnson, I. M., Alli, S. & Dragatsis, I. *PLoS Genet.* **13**, e1006846 (2017).
7. Wang, G., Liu, X., Gaertig, M. A., Li, S. & Li, X. J. *Proc. Natl Acad. Sci. USA* **113** 3359–3364 (2016).



Beverly Davidson and Alex Monteys are using gene editing to inactivate the mutated gene huntingtin.

GENE EDITING

To cut is to cure

One-off treatments that target the brain in Huntington's disease must meet strict safety and efficacy requirements.

BY MICHAEL EISENSTEIN

Beverly Davidson is well insulated from the hype of gene therapy, having spent decades working in the field. During that time, she has grappled with the harsh realities of turning flashy and potentially transformational technologies into clinical applications. But when she heard about the genome-editing technology CRISPR, she was instantly intrigued. "As soon as those first papers came out, we started playing with it," says Davidson, a specialist in neurodegenerative disease at the Children's Hospital of Philadelphia in Pennsylvania.

Like most other neurological disorders, Huntington's disease has proved to be a costly and frustrating target for drug developers. But it also has distinctive features that make it a good match for treatments that target genes. It arises from a mutation in a single gene that encodes the protein huntingtin, and a disease-causing copy of the gene can be readily distinguished from a normal copy by the presence of an over-long stretch of a repeated triplet of nucleotides, CAG (see page S36). Before turning to CRISPR, Davidson and her colleagues had some success

in treating animal models of Huntington's disease with RNA interference (RNAi), which uses synthetic molecules of RNA to prevent the production of mutant huntingtin — although it took them a considerable amount of time to get there. "We've focused the last 17 years on RNAi-based approaches," says Davidson. However, both this and a promising related treatment for Huntington's disease that involves antisense oligonucleotides (S39) will probably require long-term, repeated administration to provide sustained benefits.

By contrast, CRISPR could achieve the same benefits through a single dose that permanently inactivates the defective gene with remarkable efficiency, as Davidson's team demonstrated last year¹, both in cells from people with Huntington's disease and in mouse models of the condition. "I was surprised how easy it was — I think that's the beauty of the system," she says. In the past five years, several teams of researchers have independently shown that genome editing can reliably eliminate the gene that encodes mutant huntingtin, thereby halting the production of the toxic protein and its accumulation into clumps in experimental models.

But clearing protein clumps in mice is of questionable value when researchers often struggle to translate such findings into treatments for people — in general, potential therapies for brain disorders have a long history of failure and disappointment in clinical trials. Accordingly, the early adopters of CRISPR are trying to obtain clearer evidence of its probable clinical benefits while grappling with thorny questions related to its safety, efficacy and delivery that it is crucial to answer before trials in people can take place. "I believe we can now seriously consider clinical strategies to edit huntingtin," says Nicole Déglon, a neurologist at the Lausanne University Hospital in Switzerland, "but I would say we are still at the very beginning of the story."

TO THE LETTER

The targeted DNA-snipping capabilities of CRISPR evolved in bacteria as a defence against viruses that shoehorn their genomic material into their microbial hosts. The system uses a short sequence of RNA known as a guide RNA, which can pair with a complementary DNA sequence. Researchers have learnt how to target almost any genomic sequence by engineering an appropriate guide RNA. They couple it with an enzyme called Cas9, which can then cut both strands of a DNA sequence of interest at a specific site. Because the DNA-repair mechanism of cells is sloppy, it typically produces insertions or deletions that inactivate the affected gene.

One of the first decisions that would-be editors have to make is whether to eliminate the gene that encodes huntingtin altogether, or to selectively target the repeat-laden mutated copy. Although the function of huntingtin remains poorly understood (S36), it is crucial for early development. "If you knock out huntingtin in mice, they die in the womb," says Jong-Min Lee, a neurogeneticist at Massachusetts General Hospital in Boston. However, Xiao-Jiang Li and colleagues at Emory University School of Medicine in Atlanta, Georgia, have obtained evidence from mouse studies² that the depletion of huntingtin in the brain might not be detrimental when it occurs in adulthood. His team subsequently demonstrated³ that a CRISPR–Cas9 approach that eliminates huntingtin can clear clumps of the protein from the brain with no apparent adverse effects in a mouse model of Huntington's disease (see 'Cutting down on huntingtin'), although he is cautious about drawing too firm a conclusion. "We didn't find any obvious phenotype or neuropathology, but we still don't know whether there was some sort of functional impact," says Li.

Most researchers are therefore erring on the side of caution by designing guide RNAs that recognize sequences found only in the mutated gene. This was the approach that Davidson's team pursued, and Lee and colleagues also showed that they could make edits with remarkable accuracy in cells that were collected from a person with Huntington's disease, by designing guide RNAs that recognize sequence

variations found only on the chromosome that contains the mutated gene⁴. “The specificity is excellent,” says Lee, noting that the chromosome that bears the normal copy of the gene is consistently unaffected in treated cells. Achieving this in a clinical setting would require a level of personalization, but Lee has collected genomic data from more than 4,000 people with Huntington's disease and identified some informative patterns of sequence variation that are strongly associated with mutated copies of the gene that encodes huntingtin. “With just a couple of CRISPR designs, you could easily target more than 50% of patients,” he says.

Another concern is off-target editing, in which genes other than the target are modified inadvertently — with potentially disastrous consequences. Software can be used to predict probable off-target edits and to help researchers pick distinctive guide RNAs with reasonable confidence. But clinical researchers do need to consider the effects that CRISPR might have when used over the longer term. “We should not apply this to humans if we have permanent expression of Cas9 in the brain,” says Déglon. Unfortunately, most systems for getting the CRISPR machinery into the brain rely on its delivery by viral vector, which could lead to Cas9 being produced indefinitely. Over time, the enzyme might wreak irreparable genomic damage on healthy neurons. A possible solution entails using synthetic nanoparticles to facilitate the one-time delivery of the enzyme and guide RNA, although this work is still at an early stage.

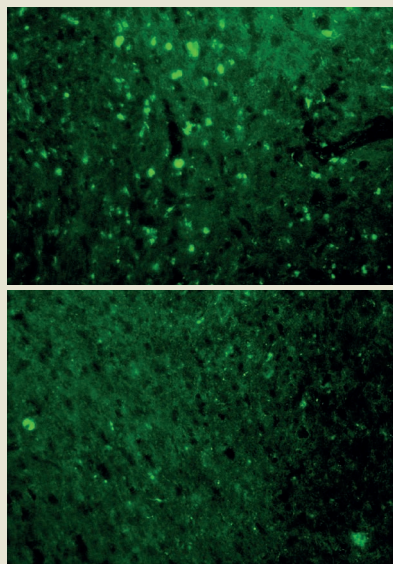
Déglon's team has devised a promising alternative to CRISPR called KamiCas9, which includes a self-destruct button for Cas9. It uses two guide RNAs — one to target the gene encoding huntingtin, and another to target the gene encoding Cas9. This means that, after a brief flurry of activity by Cas9, production of the DNA-dicing enzyme is inactivated permanently, which dramatically reduces the risk of collateral damage. She notes that several weeks after conventional CRISPR–Cas9 was applied to neural cells derived from people with Huntington's disease, low levels of off-target editing were detected — roughly 2% of modified cells received unwanted edits at a site that is particularly susceptible to off-target editing⁵. By using KamiCas9, her team was able to reduce that effect dramatically — only 0.5% of such modified cells had off-target edits. “We did not see any difference in terms of efficacy, which is really good news,” says Déglon.

BURDEN OF PROOF

Such concerns are of little relevance unless editing with CRISPR can be shown to change the course of a disease — something that is difficult to demonstrate through experiments with mice. Li's team has been able to alter Huntington's disease at the molecular level¹ by sharply reducing the production of mutant huntingtin, which forms the toxic clumps that drive the progression of the condition. “We have shown

CUTTING DOWN ON HUNTINGTIN

The mutant protein huntingtin (green fluorescence) is abundant in brain tissue gathered from a mouse model of Huntington's disease (top), but a CRISPR-based intervention that targets the gene encoding huntingtin greatly reduces production of the toxic protein (bottom).



that, in an injected area of the mouse brain, probably more than 90% of cells do not contain huntingtin aggregates,” says Li. This effect was accompanied by modest yet measurable improvements in motor function. However, as with many animal models developed for other diseases, the mouse models that researchers use to investigate Huntington's disease are poor surrogates for what happens in people with the condition. “Our model has mild motor phenotypes that show up later in life,” says Davidson. “It doesn't have any overt, robust neurodegeneration like you would see in a human patient.”

“I believe we can now seriously consider clinical strategies to edit huntingtin.”

And there are also fundamental differences in the function and organization of rodent brains compared with those of larger mammals. However, Li's team has developed a promising pig model⁶ of the condition that reflects the neurodegeneration and the motor and behavioral defects observed in people more closely than any mouse model so far. “Small animals and large animals exhibit very different pathological changes and behavioural changes,” says Li.

These improved models will also help researchers to get a handle on how many brain cells must undergo gene editing to obtain clinical gains — useful information given the impracticality and undesirability of bathing

the brain in CRISPR-laden viral vectors. The striatum, a brain structure that governs both movement and cognition, is a prominent casualty of Huntington's disease, and work with antisense oligonucleotides and RNAi suggests that efficient targeting with CRISPR could help to prevent the death of neurons in the area. Davidson thinks that cutting the production of mutant huntingtin in the striatum by half might be sufficient to halt disease progression or even prevent its onset.

For those already in the grip of Huntington's disease, there are hints that genomic repair could provide a partial rebound. “Neurons may have a lot of capacity to get rid of mutant protein if you break the continuous formation of new aggregates,” says Li. A preventive approach, however, could one day enable individuals to avert their genetic destiny, long before the onset of disease. Indeed, Huntington's disease is among the few disorders that can be confidently predicted using a genetic red flag. But even if CRISPR-based treatment amasses a strong body of preclinical data to support its use in Huntington's disease, initial clinical testing will almost certainly focus on people with symptoms, for whom improvements in motor and cognitive function can be measured in a reasonable timeframe. “Then, based on the results of the first trials showing the absence of potential side effects, they might consider early-stage or even presymptomatic patients,” says Déglon.

The brain will not be the first clinical proving ground for CRISPR. Instead, initial forays will probably be aimed at conditions such as haemophilia, which can be treated with cells that have already been genetically manipulated in the laboratory. The brain remains a daunting target because of its biological complexity, relative inaccessibility and irreplaceable function. But the parallel surge in the clinical development of gene therapy and oligonucleotide-based interventions has cleared a path for testing the potential of CRISPR in treating Huntington's disease. Even at this early stage, Davidson is optimistic. She is collaborating with Intellia Therapeutics in Cambridge, Massachusetts, which was co-founded by CRISPR pioneer Jennifer Doudna, to address the technical challenges that are involved in moving her research into the clinic. “I hate to say this, because I probably gave these sorts of numbers for RNAi, but with further advances in delivery, I could envision doing clinical testing within five years,” says Davidson. “I don't think it's particularly far off.” ■

Michael Eisenstein is a freelance science writer in Philadelphia, Pennsylvania.

- Monteys, A. M., Ebanks, S. A., Keiser, M. S. & Davidson, B. L. *Mol. Ther.* **25**, 12–23 (2017).
- Wang, G., Liu, X., Gaertig, M. A., Li, S. & Li, X. J. *Proc. Natl Acad. Sci. USA* **113**, 3359–3364 (2016).
- Yang, S. et al. *J. Clin. Invest.* **127**, 2719–2724 (2017).
- Shin, J. W. et al. *Hum. Mol. Genet.* **25**, 4566–4576 (2016).
- Merienne, N. et al. *Cell Rep.* **20**, 2980–2991 (2017).
- Yan, S. et al. *Cell* **173**, 989–1002 (2018).



Elli Hofmeister began to show signs of Huntington's disease at an early age.

PAEDIATRICS

Ahead of time

Huntington's disease is not just a condition of middle age — it can affect children and teenagers, too.

BY ELIE DOLGIN

Elli Hofmeister started to lag behind in school when she was 8 years old. By the age of 9, she needed an extra hour of tutoring each night to keep up. Elli's family chalked up her problems to a learning disability. But when Elli, at age 13, began to limp and slur her speech, "it all just started clicking," says her mother, Camille Tulenchik, a hair stylist from Maple Lake, Minnesota.

When Tulenchik was pregnant with Elli, she consulted a genetic counsellor because her boyfriend at the time had a family history of Huntington's disease. The boyfriend didn't know whether he had inherited a mutated copy of the gene huntingtin, which is responsible for the condition; if he had, there would be a 50% chance that Elli had done so, too. But if Elli did turn out to be a carrier of the gene, the counsellor explained, she probably would not develop symptoms until adulthood. Tulenchik recalls thinking, "We've got lots of time."

It was only when Elli began to experience physical problems in her early teenage years that Tulenchik decided to read up on her daughter's genetic risk. "I looked up Huntington's and saw 'juvenile,' and said, 'Oh no.'"

When nineteenth-century physician

George Huntington described the devastating neurological illness that now bears his name, he wrote that he knew of no cases in which the person affected had shown noticeable signs of disease before the age of 30. Yet the earliest documented case of juvenile Huntington's disease (JHD) pre-dates his seminal 1872 report by almost a decade — and neurologists now estimate that about 5% of cases of Huntington's disease are diagnosed before the person affected turns 20 (see 'At the extremes').

The main determinant of the age of onset is the number of repeats of a certain triplet of DNA bases in the gene huntingtin: a normal version of the gene contains 35 or fewer such repeats; 36 or more results in the formation of an unstable protein that causes Huntington's disease. The greater the number of repeats, the more unstable the protein is, and the more likely a person is to become unwell as a youngster. Elli has 65 repeats, well beyond the loosely defined threshold of 50 repeats at which JHD becomes more common. Her father has only 44 repeats, but errors in DNA replication meant that Elli inherited an even longer mutated region.

Just because someone has a large number of repeats, however, does not mean that they will show signs of Huntington's disease during their

school days. "There must be other factors that influence the onset age," says Martha Nance, medical director of the Huntington's Disease Clinic at Hennepin County Medical Center in Minneapolis, Minnesota. "We just don't know what they are."

In fact, much of JHD remains shrouded in mystery, largely because few researchers have studied the disease in young people. Take, for example, the Genetic Modifiers of Huntington's Disease Consortium, which undertook the largest DNA-mapping study of genes associated with the progression of Huntington's disease (GeM-HD Consortium, *Cell* **162**, 516–526; 2015). Of the 4,082 participants in the study, only 29 had been diagnosed before the age of 20, according to neurogeneticist Jong-Min Lee, one of the consortium's leaders at the Massachusetts General Hospital in Boston.

In recent years, researchers' interest in JHD has picked up — and slowly the spotlight is shifting to this unique population of patients. "For too long, JHD has been under the radar," says Peg Nopoulos, a psychiatrist and neuroscientist at the University of Iowa in Iowa City. "It's time to pay attention to the kids who are suffering from this disease."

CATCH THE SIGNS

For Nopoulos, filling in the missing data meant starting with a simple catalogue of the many ways in which symptoms differ between children and adults with Huntington's disease. Among young people with the condition, muscle stiffness is perhaps the most common complaint. That's because children typically develop rigidity as one of the initial movement-related symptoms, and rarely exhibit the jerky, involuntary movements known as chorea that characterize adult-onset disease. However,

ACKERMAN + GRUBER

when Nopolous and her colleagues surveyed caregivers of patients with JHD, they also learned of a range of other problems found nowhere in the medical literature.

As Nopoulos and her team reported last year (A. D. Moser *et al. Neurodegener. Dis. Manag.* 7, 307–315; 2017), more than three-quarters of the respondents said that their wards experienced tics, 69% said they were in some type of pain, and around half said they were dealing with moderate-to-severe itching. These symptoms were recorded rarely in adults, but seemed to be widespread in children with JHD. “It suggests that juvenile-onset Huntington’s disease is impacting on parts of the brain in a different way than in an adult-onset disease,” says Nance, who collaborated on the survey.

To further probe those neurological differences, Nopoulos has used magnetic resonance imaging to scan the brains of about 25 children with JHD (including Elli), as well as those of hundreds of healthy young people. A defining feature of Huntington’s disease is that nerve cells of the striatum, a motor-control region in the centre of the brain, shrivel and die as the disease progresses — and, indeed, in the study participants with JHD, “the striatum is just toast,” Nopoulos says.

However, the scans also revealed that as the striatum shrinks in these children, another movement-related brain structure — the cerebellum — gets larger. This “pathological compensation”, as Nopoulos calls it, could explain why youngsters with Huntington’s disease seem to skip the chorea stage of the condition and go straight to stiffness.

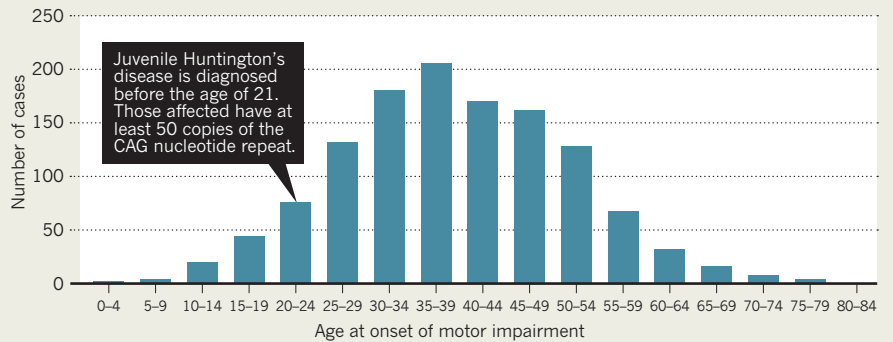
By growing too big, the cerebellum doesn’t just make up for the missing motor functions of the striatum; it overshoots the mark and puts the brakes on movement entirely.

Nopoulos presented these findings in February at the 13th Annual Huntington’s Disease Therapeutics Conference — at which one of the few other scientists to discuss data on JHD was Mahmoud Pouladi, a neurogeneticist at the A*STAR Translational Laboratory in Genetic Medicine and the National University of Singapore. Pouladi’s team coaxed stem-cell lines generated from children with Huntington’s disease to form what amount to 3D miniature brains. The disease is usually associated with neurodegeneration, but experiments with Pouladi’s brain-like structures suggest that it’s also linked to neurodevelopment — and the greater the number of triplet repeats, the more abnormal that development will be.

Another way to study the molecular basis of JHD — and to try to develop treatments to reverse the condition — is to use transgenic mouse models. Few scientists who genetically engineer mice to study Huntington’s disease set out explicitly to model JHD rather than adult-onset disease. But according to Gillian Bates, a molecular neuroscientist at University College London, that might be what the research community has done inadvertently. “All of our mouse models are models of the

AT THE EXTREMES

Motor impairment associated with Huntington’s disease is rare in children and adolescents. People who carry the gene mutation that causes the condition develop symptoms on average at around the age of 40. Huntington’s disease can strike later in life, but this is also rare.



juvenile form of the disease,” she says.

To observe neurodegeneration during the short life of a mouse — and over a time course that’s suitable for experimentation — “we often purposefully push the disease”, explains Cat Lutz, director of the mouse repository at the Jackson Laboratory in Bar Harbor, Maine. For Huntington’s disease, that means increasing the number of triplet repeats to a level that would cause childhood onset in people.

This protocol could explain why most mouse models show many hallmarks of JHD, including rigidity and susceptibility to seizures — and might even call into question the validity of extrapolating data from mice to adult-onset Huntington’s disease. It could also mean that scientists know more about the basic neurology of JHD than they realize.

Then again, those symptoms could just be a reflection of how Huntington’s disease manifests in a rodent, and might have nothing to do with the number of triplet repeats or types of the disease in people. The truth, says David Howland, director of research on new animal models for Huntington’s disease at the CHDI Foundation, a US non-profit organization, is that “we don’t know how good our models really are”.

MATTER OF SCALE

More effort is being invested in developing tools for the clinical investigation of JHD. A working group of the European Huntington’s Disease Network, led by clinical geneticist Oliver Quarrell at Sheffield Children’s Hospital, UK, ran a five-year observational study that tracked 95 people who had been diagnosed with Huntington’s disease at or before the age of 25 using the Unified Huntington’s Disease Rating Scale, the most widely used and best-validated metric of clinical progression (see page S46).

The results are not yet published, but Quarrell says that the evaluation tool was unsuitable for measuring motor functions in these young patients because it puts great emphasis on chorea and much less on symptoms related to rigidity. He and his colleagues are now working on a modified scale to better match the distinct features of JHD.

That tool will be important in light of a ruling by the European Medicines Agency stating that, from July 2018, companies that develop drugs for Huntington’s disease will have to test such treatments in paediatric populations before the products can receive marketing approval. At present, all drugs used to manage the symptoms of JHD — including dopamine modulators, anti-seizure medications, anti-anxiety agents and muscle relaxants — are taken off-label. Elli, for instance, uses a drug that is commonly prescribed for Parkinson’s disease to ease her stiffness, over-the-counter pain medicines to deal with aches, and physical therapy to stay as supple as possible.

Her mother follows websites such as HDBuzz to keep on top of the latest drug trials. She then discusses options with Nance, Elli’s neurologist, but is yet to find anything promising that also accepts younger participants. To enrol in a study for one of the treatments that aims to silence the mutated gene huntingtin (S39), for example, volunteers need to be at least 25 years old. “Right now I feel like we are very limited in our options,” says Tulenchik.

Elli turned 20 in February. Three days a week, she attends a transition programme for young adults with special needs, where she helps to run the coffee shop. She also volunteers at a nearby nursing home, decorating the bulletin board and cleaning bingo cards, swimming-pool noodles and musical instruments. For her most recent birthday, Elli celebrated by hosting a sleepover for only her closest female relatives, including her sister Violet, which meant that her brother Zander couldn’t attend. “No boys allowed!” says Elli, in a slow and indistinct manner.

They decorated masks, ate cake and ice cream, and stayed up after midnight, watching *Fly Away Home*, a feel-good, 1990s-era family drama about a teenager who teaches her pet geese to fly. “Our motto is: ‘Today is our best day,’” Tulenchik says. “We just focus on today.” ■

Elie Dolgin is a science writer in Somerville, Massachusetts.



CLINICAL TRIALS

The endpoint is near

Better measures of efficacy are needed in trials of treatments for Huntington's disease.

BY KAT ARNEY

Potential treatments for Huntington's disease are starting to pass through clinical trials, and excitement is building among researchers. "I've been working in Huntington's disease for more than 20 years, and we're in a new era," says Sarah Tabrizi, a clinical neurologist at University College London. "We've previously only had trials looking at drugs that relieve symptoms, but now we know the root cause of the disease, and we're starting to see molecular therapies that target it," she says.

Tabrizi led one such study, which tested whether short pieces of modified DNA known as antisense oligonucleotides (ASOs) could switch off the protein-coding messages transcribed from the gene that is mutated in people with Huntington's disease. The trial received widespread media coverage in December 2017, and researchers, clinicians and patients hope that this gene-silencing approach could provide the first treatment to truly modify the disease (see page S39).

But beneath the positivity lurks a thorny issue for researchers, such as Tabrizi, who are developing treatments. "We absolutely have to be sure they are working," she says. Unfortunately, this is not so easy to determine in Huntington's disease.

Unlike trials of cancer drugs, in which

efficacy can be quantified with relative ease — tumours can be seen to shrink, grow or stay the same size — designing trials to demonstrate meaningful improvements in progressive neurological conditions such as Huntington's disease is not straightforward.

The assessment tool used in almost all trials so far is the Unified Huntington's Disease Rating Scale (UHDRS). Developed in 1996 by the Huntington Study Group, an international collaboration, the UHDRS enables doctors to score a person's overall physical and neurological fitness. By testing participants at regular intervals, investigators can work out whether a potential treatment is slowing the progression of the disease, compared with a placebo.

"We've used the UHDRS for many years," says Blair Leavitt, consulting neurologist at the University of British Columbia in Vancouver, Canada. "You can see reliable progression over a certain period, but we need better ways to measure it."

Symptoms such as movement difficulties or cognitive impairment can vary in severity from day to day. This makes it difficult to tell whether a treatment is having an impact on the disease or the patient is just having a good day. And, as Leavitt explains, the subjective nature of functional assessments such as the UHDRS means that they're greatly susceptible to the power of the placebo effect.

The reliability of such tools was thrown into the spotlight with the announcement of results from Pride-HD, a trial of pridopidine (Huntexil) run by Teva Pharmaceutical Industries in Petach Tikva, Israel. Although previous trials of drugs aimed at relieving symptoms were inconclusive, preclinical testing suggested that pridopidine might help to protect neurons from the damaging effects of mutant huntingtin, the toxic protein produced by people with Huntington's disease.

Yet Pride-HD failed to show that pridopidine led to improvements in motor function, which had been declared as the study's primary endpoint (a predetermined milestone that signals the success of a treatment). However, the researchers did notice some improvement in one of the six components of the UHDRS, a measure known as total functional capacity. As one of the most subjective parts of the scale, it considers whether a person is able to work, handle finances or perform self-care tasks.

PLACEBO EFFECT

The results generated hope that pridopidine might modify the progression of Huntington's disease, as opposed to its symptoms. But, as Leavitt points out, it's more likely that there is an alternative explanation for its effect.

"What's pretty clear to me is that there's a big placebo effect seen with investigator-rated

NEIL WEBB

scales like UHDRS that isn't there in more quantitative measures," he says. "This is why you must define your primary endpoint before the trial — you can't keep going back and looking until you find something that works."

In the absence of better options, and despite its limitations, the UHDRS has been the primary endpoint of choice for two decades. "Objective, quantitative measures will give us more sensitivity," says Leavitt. In 2012, Tabrizi, Leavitt and their collaborators published the results¹ of TRACK-HD — a major study that used a battery of approaches, including brain imaging and cognitive, motor and psychiatric assessments, to monitor more than 100 people with early-stage Huntington's disease over a period of two years. The study also followed carriers of the mutated gene who were yet to show signs of the disease, as well as people without the mutation.

They found a number of measurable features that reflected disease progression, including brain volume on magnetic resonance imaging scans and specific motor and cognitive characteristics. The TRACK-HD team have pooled their data with results from other cohort studies to generate a composite endpoint for future trials. It comprises a suite of cognitive, motor and physiological traits that can accurately assess the progression of Huntington's disease.

Known as the composite UHDRS, it's built on the bones of the original scale and retains the most reliable tests. It also includes further cognitive measures and ditched traits that don't show much progression with time, such as emotional recognition and tongue-muscle strength.

The combined study, which included more than 1,600 people with early-stage Huntington's disease, defined important parameters such as the number of participants that is needed to ensure statistical rigour, as well as the optimal duration of a trial. "Huntington's is a slowly progressive disease, but we showed that we could measure progression in almost everybody over just two years," says Tabrizi.

The team also reached agreement on the degree

"We have exciting new measures and therapies, but we don't have a good way of comparing them."

to which progression should be slowed for trials of disease-modifying treatments to be deemed a success: people who receive the treatment should show a decline on the composite UHDRS that is 20–30% slower than that of those who receive a placebo.

It seems obvious that trials should be designed to paint the most accurate picture of the benefits and risks of treatments. But Tabrizi suggests that the lack of effective disease-modifying treatments for Huntington's disease, together with the fact that drugs such as ASOs must be administered directly into cerebrospinal fluid (CSF) by lumbar puncture — an uncomfortable procedure in which a needle is inserted into the spinal canal — means that

researchers have an ethical duty to make sure that trials are designed as well as possible to reveal whether treatments are working.

MARKERS OF SUCCESS

Although Leavitt and Tabrizi agree that the composite UHDRS is an improvement on the conventional scale, the hunt is on for biological markers (biomarkers) that change as Huntington's disease progresses. The most obvious candidate is mutant huntingtin, which was thought to leach into CSF from damaged brain cells, in a similar way to CSF biomarkers now used for other neurodegenerative conditions. However, developing a reliable test for quantifying the protein has proved challenging, in part because it is present in CSF at very low concentrations.

As a solution, Leavitt and his colleagues are developing ultrasensitive assays that can detect changes in levels of huntingtin in CSF with disease progression. One approach², based on immunoprecipitation and flow cytometry, revealed a decrease in the mutant protein following treatment with ASOs in a mouse model of Huntington's disease, and could serve as a primary endpoint for trials of disease-modifying treatments. An alternative approach³ that counts single molecules of mutant huntingtin was used in a 2015 trial of ASOs (S39). In any case, monitoring protein levels in CSF would require repeated invasive lumbar punctures.

Other teams are using blood as a more easy-to-access source of potential biomarkers. One such candidate molecule is neurofilament light polypeptide (NF-L) — a component of neurons that is released into CSF as the cells die, eventually making its way into blood. Levels of NF-L in blood plasma mirror those of mutant huntingtin in CSF, and a retrospective study⁴ of more than 200 people with Huntington's disease or who carry the mutation that leads to the condition showed that NF-L levels could be used to predict the onset of symptoms, as well as to track disease progression.

Despite promising results, biomarkers in blood or CSF are only surrogates for the underlying disease that ravages the brain. The most direct assay would involve imaging mutant huntingtin in the brain to determine whether it diminishes after treatment, although this is technically challenging. Researchers funded by the US non-profit CHDI Foundation are developing radioactive 'flags', or ligands, that bind to clumps of mutant huntingtin in the brain and can be detected by positron emission tomography. Trials in people are expected to start later this year, according to Cristina Sampaio, chief medical officer at CHDI.

Leavitt and his team are also interested in using wearable sensors to monitor certain biomarkers such as changes in gait or cognitive function in real time. Investigators who use the UHDRS or similar scales can assess the abilities of patients only on the days on which they visit the clinic. However, sensors such as accelerometers can take measurements continually over periods of days, weeks or months, and are even

able to monitor sleep patterns and activity levels.

This data stream can be relayed from people's homes to the clinic, creating a more-detailed profile of symptom progression. Initial studies⁵ used lab-improvised systems of sensors that are strapped to the chest, wrist and ankles. But off-the-shelf technologies such as fitness trackers or smart watches, in conjunction with smartphones, are likely to become a more practical option. "We can design simple tests on a smartphone to measure gait, walking speed or cognitive function, and we can collect daily data on mood or any other problems they might be having," Leavitt says.

Real-time, remote monitoring of such biomarkers through smartphone apps could reduce the burden of taking part in trials for participants and carers. Travelling long distances to a hospital or trial centre can be arduous, especially for people with advanced Huntington's disease. And the continuous collection of data would make it easier for researchers to follow overall disease-progression trends and to build a more accurate idea of each person's response to treatment.

Despite progress being made, those who are developing new primary endpoints find themselves in a chicken-and-egg situation. To show that they work, trials of disease-modifying treatments need more-appropriate endpoints than those provided by the UHDRS. But the improved endpoints can be validated only against effective drugs, to demonstrate that they accurately measure disease progression and patients' responses to treatment. The current generation of trials is beginning to incorporate measures such as biomarkers and brain imaging as exploratory secondary endpoints, alongside the UHDRS. Despite its flaws, the UHDRS is still the only tried-and-true measure of disease progression available to researchers.

"It's a circular problem. We have exciting new measures and therapies, but we don't have a good way of comparing them to prove that they work," says Leavitt. "Our main clinical endpoint is still the old UHDRS, which isn't that great. We're at the point now where we need an effective therapy to show how things respond."

As Huntington's disease enters an era of targeted molecular treatments, Tabrizi thinks that researchers owe it to those affected to design the best possible trials in which to test such drugs. "We've spent years studying the natural history of the disease to develop our armamentarium for these trials, and we're just waiting for really good drugs," she says. "Huntington's is a terrible disease with a huge unmet need, and patients and their families desperately want treatments that work. We cannot afford to mess this up." ■

Kat Arney is a science writer and broadcaster based near London.

1. Tabrizi, S. J. *et al. Lancet Neurol.* **11**, 42–53 (2012).
2. Southwell, A. L. *et al. Sci. Rep.* **5**, 12166 (2015).
3. Wild, E. J. *et al. J. Clin. Invest.* **125**, 1979–1986 (2015).
4. Byrne, L. M. *et al. Lancet Neurol.* **16**, 601–609 (2017).
5. Andrzejewski, K. L. *et al. J. Huntingtons Dis.* **5**, 199–206 (2016).

HUNTINGTON'S DISEASE

4 BIG QUESTIONS

Although potential treatments are now entering the pipeline, the molecular cause and progression of Huntington's disease continue to elude researchers.

BY ANNA NOWOGRODZKI

QUESTION

WHY IT MATTERS

WHAT WE KNOW

NEXT STEPS

1

How does the mutant protein huntingtin cause Huntington's disease?

Huntington's disease is caused by a mutation in a single gene called huntingtin (*HTT*), which encodes the protein huntingtin. Understanding how the mutant protein causes disease could open up avenues for treating the condition and its symptoms.

Mutant huntingtin forms clumps inside the cell that seem to interfere with communication along the axons of neurons. Such aggregates can also throw a wrench into the transcription of other genes and hinder cells' waste-removal systems.

Various projects led by researchers, companies and non-profit organizations are using computational methods to better understand the shape of mutant huntingtin, how it aggregates, and how it interacts with other proteins in the cell.

2

What is the role of normal huntingtin?

To help develop treatments for Huntington's disease, including those that use RNA interference, antisense oligonucleotides (ASOs) or gene editing, it's important to determine whether normal huntingtin, as well as the mutant version, can be reduced or eliminated safely.

Blocking *Htt* expression in mouse embryos is lethal. In adult mice, some studies show that removing normal huntingtin has only limited effects, whereas others indicate it shortens lifespan and causes nerve and behavioural problems. The effect of reducing huntingtin in people is unknown.

Researchers are eliminating normal huntingtin in mammals with lifespans longer than those of mice to determine any long-term effects. Efforts are also underway to inactivate just the mutated copy of *HTT*, leaving the normal version intact, using the gene-editing tool CRISPR-Cas9.

3

How can we better characterize the progression of Huntington's disease?

To more effectively assess treatments in trials, doctors need improved ways of measuring whether they slow disease progression. The current best tool is the clinician-rated Unified Huntington's Disease Rating Scale, which is reliable but prone to the power of the placebo effect.

A 2017 study showed that changes in levels of neurofilament light polypeptide (NF-L) in blood correlate with the onset of Huntington's disease, making it a possible biomarker. Other biomarkers that correlate with the condition can be measured by functional brain imaging.

Three large long-term observational studies have been designed to assess the ability of potential biomarkers to measure disease progression. The team investigating NF-L has launched a 600-participant study, and is already monitoring NF-L levels in at least 80 people.

4

Will ASOs be the first effective treatment for Huntington's disease?

ASOs are the first potential treatment to have successfully lowered levels of mutant huntingtin in trials conducted in people. But it's uncertain whether these molecules can slow or halt progression of Huntington's disease.

In a phase I/IIa trial, an ASO called IONIS-HTT_{Rx} reduced the levels of mutant huntingtin in participants' cerebrospinal fluid. But the trial was too short to determine the treatment's long-term effects. The drug is delivered once a month via an injection into the spine.

Further trials of IONIS-HTT_{Rx} with larger numbers of participants are needed to determine whether the drug is effective at treating Huntington's disease. Researchers are also monitoring the 46 participants of the initial trial for any long-term effects.

Anna Nowogrodzki is a freelance science writer based near Boston, Massachusetts.



Visitors watch a robot during the 2015 China Yiwu International Manufacturing Equipment Expo & China Intelligent Expo in Yiwu, Zhejiang province.

Zhejiang province is open for science business

In the global race for high-tech excellence, a historic region of China is taking a prominent role on the international stage.

BY SARAH O'MEARA

Every weekend, busloads of tourists appear at the newly opened International Campus Zhejiang University in Haining City. It is a 20-minute high-speed train ride away from the province's historic capital city of Hangzhou, home to the campus's prestigious parent institution, Zhejiang University. The 80-hectare plot, with its modern architecture, water features and pleasant walkways, is a dream for day trippers, who come to picnic by the artificial lake.

The arrival of the tourists initially came as a surprise to Philip Krein, full-time dean of the Zhejiang University/University of Illinois at Urbana-Champaign Institute in Haining. Krein moved there from the United States in 2016 and

was not used to seeing universities as tourist destinations. But he's getting used to the flag-bearing tour guides ushering visitors around the facilities, which opened in 2017. His is one of a number of overseas universities that have daughter institutions or collaborative labs in Zhejiang, including Imperial College London and the University of Edinburgh, UK.

"Academic institutions are extremely important and valued in China," he says. "People want to see how the country is developing."

Although the region's historic strengths lie in its prominence as a shipping hub and as an access point to central China, Zhejiang's government leaders think that future economic growth will come from investment in

its digital economy. At every level of public life, from university programmes to city management, officials are working with scientists and engineers to put cutting-edge science — such as artificial intelligence, big data and cloud computing — at the heart of the region's development, and to further internationalize the area. Zhejiang's global significance was affirmed in 2016, when Hangzhou hosted the first Group of 20 (G20) meeting of world leaders ever to be held in China.

The economic change taking place in Zhejiang reflects the country's wider ambitions (see 'On the map'). China's economy is in a period of rapid transition. The government's goal, originally set out in a 15-year science ►

► and technology plan in 2006, is to transform the country from a low-cost manufacturer into a technologically advanced, innovative economy in which products are no longer just made, but “created”, in China.

By 2049 — the centenary of the founding of the People's Republic of China — the country plans to be a world-leading science and technology power. It also wants to share its expertise with trading countries across Asia, Europe, the Middle East and East Africa through higher-education initiatives, as part of its ‘One Belt, One Road’ foreign-policy programme, initiated in 2013. That project aims to strengthen economic and diplomatic relationships between China and regional trading partners through the construction of vast infrastructure projects, including roads, railways and docks.

BIG PICTURE

Krein says that the clear focus of Zhejiang's government and its international outlook enables faculty members at his university to think big. On campus, the three foreign universities are in the early stages of planning collaborative projects, such as one at the Biomedical Translational Research Centre, a facility announced in 2018 to focus on turning academic research into technologies that can improve health care.

Krein's team is also developing a large-scale ‘intelligent’ infrastructure project that uses millions of sensors to evaluate the distribution of bridges, roads and railways, alongside sensors in the water systems, to build up a comprehensive picture of the flow of traffic, waste and people across urban areas. Both projects are in their early stages, but Krein expects funding to come from a consortium of industry and local government.

“We're not a satellite organization. This is a genuine collaboration,” explains Krein. He is trying to develop unusual, cross-discipline projects and foster teaching styles that encourage greater creativity and curiosity among Chinese students — an endeavour spurred by a wider concern that China's teaching culture, from an early age, prioritizes rote learning over critical thinking.

A decade ago, China's government launched an initiative to reverse its academic brain drain, after research revealed that of 1.1 million Chinese people who had left to study overseas since 1978, only 275,000 had returned, as of June 2007 (see go.nature.com/bdrain). High salaries were offered by universities to Chinese-born, Western-trained professors willing to return to China as part of a programme called 1000 Talents.

Despite such efforts, it's still common for ambitious young scientists to head abroad after finishing their degree in China, because experience in foreign labs remains highly valued. This situation can make it difficult for Chinese universities to recruit high-level talent at the postdoc level, says Jiaming Hu, a neuroscientist at Zhejiang University. Hu says that universities often require newly hired

ON THE MAP

Zhejiang province is less well known than other Chinese regions, but its health is vital to the country's rapidly internationalizing science economy.



professors and associate professors to have at least two years of research experience abroad.

Yet early-career Chinese scientists who choose to stay in Zhejiang can be in a strong position to find rewarding jobs, both in state-funded and commercial labs, given that the province is betting much of its future on science and technology development. Hu's lab is just 3 years old, and the product of a US\$25-million, 5-year investment into the institute. That money also bought a large primate facility, which Hu says is unlike anything he'd find in the West. “Some of my friends went to foreign labs and sometimes find it not that easy to get all the animal resources that they need,” Hu says.

Researchers are also finding that, even if they stay in China, opportunities to collaborate with foreign scientists are growing. Materials scientist Yong Wang, who returned to China in 2012 as part of the 1000 Talents programme, says that Zhejiang University offers multiple programmes to enable staff and students to go abroad for a short stint.

Guohua Xu, an evolutionary geneticist who works alongside Jiaming Hu at Zhejiang University, adds that leading scholars frequently visit the campus and give talks. Both Hu and Xu highlight the generous funding researchers in China enjoy. “The advantage is that we often have equipment many Western universities do not have,” Xu says. “The disadvantage is that we are not yet at a high enough level to compete with the best international teams and attract the best talent.”

But that generosity isn't always balanced across research fields. A reform of the funding system, announced in March, seems to shift priorities away from basic-science research and towards government-sanctioned projects, at a time when many prominent Chinese researchers say that China's spending is already too low compared with other nations.

Wang remains optimistic. “I think it's possible to do both. To work within international teams and pursue universal truths, and also achieve China-specific goals,” he says.

For entrepreneurial scientists in Zhejiang, many routes are available for turning research

into commercial products. Zhejiang University has invested in its own industrial park, 1 of 15 similar science parks across China that are designed to act as incubators for commercial enterprises and start-ups. Funding comes from both the private and public sectors. Recent commercial successes include Dore Technology, which offers tourism-focused smart technology such as interactive maps and intelligent audio guides, and NationalChip, which manufactures computer chips for television systems.

Last year, the province's capital opened a 113-square-kilometre zone for the development of science and technology, called Hangzhou Future Sci-TechCity. Within the site lies AI Town. It opened last year, and plans to have 20,000 researchers and 200 innovation teams, led by top scientists and industry pioneers, on site by 2022. The vast collaboration will bring together innovative companies, such as Alibaba and the Chinese Internet giant Baidu, with a variety of prominent university teams.

The impetus to turn research into viable products is embedded into university culture, says Anna Wang Roe, a neuroscientist at Zhejiang University. Staff and postdocs are encouraged to apply for opportunities to set up new ventures. All of this comes with a price tag: each year, the province invests around \$20 million to \$30 million in Zhejiang University to encourage the development of companies. In early 2018, the province reported that, between 2013 and 2017, the number of high-tech enterprises had more than doubled, to 11,462, whereas small and medium-sized science and technology enterprises had increased eightfold, reaching 40,440.

PIONEERING ROOTS

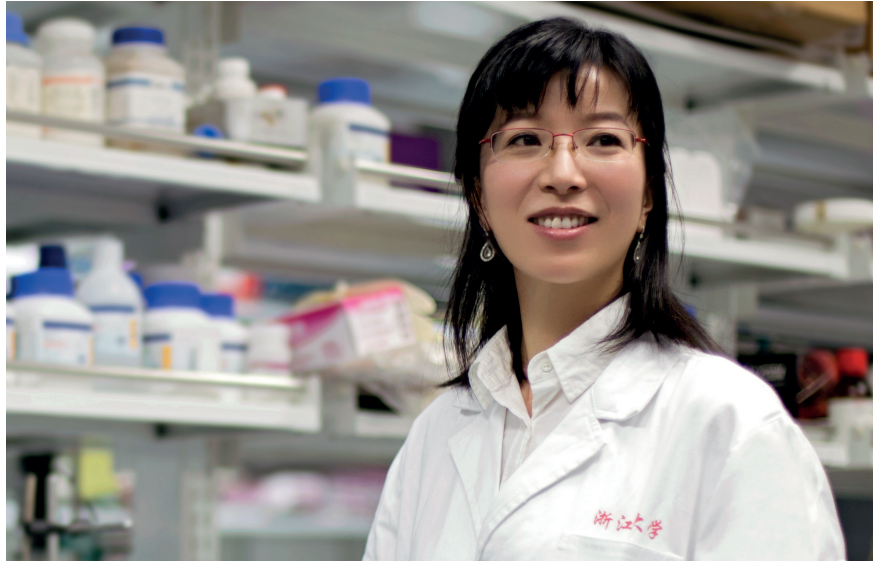
“The local government here has always looked ahead,” says May Tan-Mullins, who studies international relations at the University of Nottingham Ningbo China. “Fifteen years ago, as the city was growing, they planted trees, when no other Chinese cities did. Locals complained they were wasting money that should be spent on houses. Now people appreciate the city's green belt.”

Tan-Mullins's institution — a joint venture between the University of Nottingham, UK, and the Zhejiang Wanli Education Group — is a reflection of an internationalizing China. “We have staff from over 50 countries and students from more than 70,” says Tan-Mullins.

Although the university's major market is still domestic, about 10% of students come from overseas — the highest proportion of any Chinese university. And 90% of faculty members are international. “We joke that on our 144 acres of land, we probably have the highest concentration of foreigners in China,” says Tan-Mullins.

“It's brought Ningbo on to the global stage,” she says. “And also onto the China stage.” ■

Sarah O'Meara is a writer in Shanghai. Additional research by Liu Shaoxin.



Q&A Hailan Hu

The zest of Zhejiang

Neuroscientist Hailan Hu moved back to her home town to capitalize on its vibrant science environment. She researches depression at Zhejiang University School of Medicine in Hangzhou.

Why did you return to Hangzhou?

The medical school has a supportive and nurturing environment where I could grow. Being close to family and friends is a plus.

What were you doing before that?

After graduating from Peking University in Beijing, I did a PhD at the University of California, Berkeley, followed by a postdoc at Cold Spring Harbor Laboratory in New York. I joined the Chinese Academy of Sciences Institute of Neuroscience in Shanghai in 2008.

What do you like about Zhejiang University?

We foster a strong culture of interdisciplinary research. My own lab collaborates with engineering, computer science, pharmacology and chemistry teams. Zhejiang University has several affiliated hospitals that provide a good platform for translational medical research, and it has a reputation for a strong entrepreneurial spirit.

Tell us about your team and current research.

It's an entirely Chinese team right now, but I'm considering taking on foreign students as we grow. In 2016, we discovered how the anaesthetic ketamine blocks electrical bursts from a region of the brain and relieves the symptoms of severe depression. We're talking to scientists and clinicians worldwide about translating the research into antidepressants.

Are you planning new collaborations abroad?

Our neuroscience centre is going to establish a programme with the University of Toronto in Canada, for students and postdocs to work overseas. We have formal collaborations with the University of California, Los Angeles, Columbia University in New York City and the University of Melbourne in Australia.

How has Hangzhou changed since you left?

Hangzhou was China's top honeymoon destination. Now, it is still beautiful, but its tourism-based economy has transformed. The atmosphere is modern and international. It feels different. I almost never take my wallet out; I pay using my mobile phone.

Are you surprised by the transformation?

Change was always a part of life here. There's a saying: the three big eastern economies, Beijing, Shanghai and Zhejiang, have different business models. Beijing has the government-owned businesses, Shanghai favours foreign companies and big brands and Zhejiang cultivates entrepreneurship.

What practical benefits do you find China has?

I perhaps spend less time writing grants compared with my peers in the United States.

What advances do you hope to see in the Chinese research environment in the future?

As basic research booms and more graduate students and postdocs are doing outstanding work in China, I hope we can provide equal support to those trained at home and abroad. At present, some career-development grants are designed for Chinese scientists who have studied and worked abroad. Now is the time to extend these schemes to all qualified young trainees.

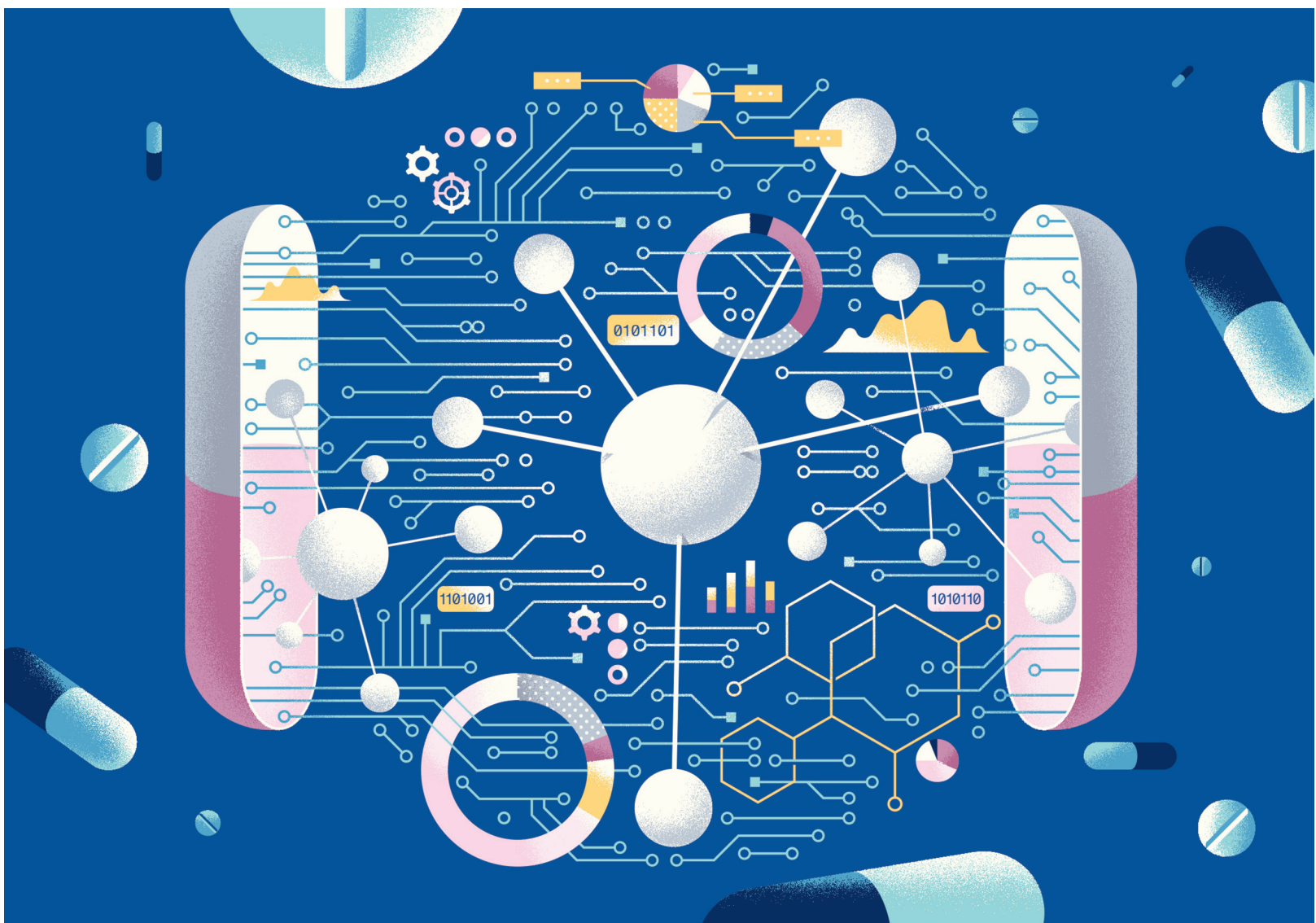
What attracts young scientists to Zhejiang?

With its nice environment and low cost of living, I think Zhejiang has become as attractive to researchers as Beijing or Shanghai, if not more so. The opportunities are growing. ■

INTERVIEW BY SARAH O'MEARA.

ADDITIONAL RESEARCH BY LIU SHAOXIN

This interview has been edited for length and clarity.



Computer-calculated compounds

Researchers are deploying artificial intelligence to discover drugs.

BY NIC FLEMING

An enormous figure looms over scientists searching for new drugs: the estimated US\$2.6-billion price tag of developing a treatment. A lot of that effectively goes down the drain, because it includes money spent on the nine out of ten candidate therapies that fail somewhere between phase I trials and regulatory approval. Few people in the field doubt the need to do things differently.

Leading biopharmaceutical companies believe a solution is at hand. Pfizer is using IBM Watson, a system that uses machine learning, to

power its search for immuno-oncology drugs. Sanofi has signed a deal to use UK start-up Exscientia's artificial-intelligence (AI) platform to hunt for metabolic-disease therapies, and Roche subsidiary Genentech is using an AI system from GNS Healthcare in Cambridge, Massachusetts, to help drive the multinational company's search for cancer treatments. Most sizeable biopharma players have similar collaborations or internal programmes.

If the proponents of these techniques are right, AI and machine learning will usher in

an era of quicker, cheaper and more-effective drug discovery. Some are sceptical, but most experts do expect these tools to become increasingly important. This shift presents both challenges and opportunities for scientists, especially when the techniques are combined with automation (see 'Here come the robots'). Early-career researchers, in particular, need to get to grips with what AI can do and how best to acquire the skills they need to be employable in the job market of tomorrow.

The AI pioneers of the 1950s discussed ►

► building machines that could sense, reason and think like people — a concept known as ‘general AI’ that is likely to remain in the realms of science fiction for some time. However, the continued rapid growth in computer-processing power over the past two decades, the availability of large data sets and the development of advanced algorithms have driven major improvements in machine learning. This has helped to bring about ‘narrow AI’, which focuses on specific tasks. These include improved abilities to analyse, understand and generate text and speech through an AI technique called natural-language processing, and artificial neural networks designed to mimic the way our brains make sense of the world. Such techniques are already in widespread use in fields such as computer vision, voice analysis and route selection. This progress has also triggered a wave of start-ups that employ AI for drug discovery, with many of them using it to identify patterns hidden in large volumes of data.

For example, researchers at biotechnology company Berg, near Boston, Massachusetts, have developed a model to identify previously unknown cancer mechanisms using tests on more than 1,000 cancerous and healthy human cell samples. They modelled diseased human cells by varying the levels of sugar and oxygen the cells were exposed to, and then tracked their lipid, metabolite, enzyme and protein profiles. The group uses its AI platform to generate and analyse immense amounts of biological and outcomes data from patients to highlight key differences between diseased and healthy cells.

The aim of Berg’s approach is to identify potential treatments on the basis of the precise biological causes of disease. “We are turning the drug-discovery paradigm upside down by using patient-driven biology and data to derive more-predictive hypotheses, rather than the traditional trial-and-error approach,” says Niven Narain, Berg’s co-founder and chief executive.

Using this approach, Narain’s team identified the importance of certain naturally occurring molecules in cancer metabolism. This led the group to discover how a new cancer drug works, and indicated some possible therapeutic uses. The drug, BPM31510, is currently in a phase II clinical trial involving people with advanced pancreatic cancer. The company is also using this AI system to look for drug targets and therapies for other conditions, including diabetes and Parkinson’s disease.

London-based start-up firm BenevolentBio has its own AI platform, into which it feeds data from sources such as research papers, patents, clinical trials and patient records. This forms a representation, based in the cloud, of more than one billion known and inferred relationships between biological entities such as genes, symptoms, diseases, proteins, tissues, species and candidate drugs. This can be queried rather like a search engine, to produce ‘knowledge graphs’ of, for example, a medical condition and the genes that

are associated with it, or the compounds that have been shown to affect it. Most of the data that the platform crunches are not annotated, so it uses natural-language processing to recognize entities and understand their links to other things. “AI can put all this data in context and surface the most salient information for drug-discovery scientists,” says Jackie Hunter, chief executive of BenevolentBio.

“AI IS GOING TO LEAD TO THE FULL UNDERSTANDING OF HUMAN BIOLOGY AND GIVE US THE MEANS TO FULLY ADDRESS HUMAN DISEASE.”

When the company asked this system to suggest new ways to treat amyotrophic lateral sclerosis (ALS), also known as motor neuron disease (MND), it flagged around 100 existing compounds as having potential. From these, scientists at BenevolentBio selected five to undergo tests using patient-derived cells at the Sheffield Institute of Translational Neuroscience, UK. The research, presented at the International Symposium on ALS/MND in Boston, Massachusetts, in December 2017, found that four of these compounds had promise, and one was shown to delay neurological symptoms in mice.

PATTERN RECOGNITION

Despite these promising applications, many scientists are unaware of the capabilities of AI. A survey published in February by BenchSci, a start-up in Toronto, Canada, that provides a machine-learning tool for scientists searching for antibodies, found that 41% of the 330 drug-discovery researchers who took part were unfamiliar with the uses of AI (see go.nature.com/2xarpt3).

Leaders in the field think that researchers should brush up on this knowledge as soon as possible.

“AI is going to lead to the full understanding of human biology and give us the means to fully address human disease,” says Thomas Chittenden, who leads a team at Wuxi NextCODE in Cambridge, Massachusetts. Wuxi NextCODE was formed in 2015 after drug-discovery firm WuXi AppTec in Shanghai, China, acquired NextCODE Health, a spin-off from Icelandic company deCODE Genetics. “The way we develop drugs and assess them in clinical trials will all come down to very sophisticated pattern recognition,” he says.

In May 2017, a group including researchers

at Yale University in New Haven, Connecticut, demonstrated the role of a family of proteins called fibroblast growth factors (FGFs) in blood-vessel development (P. Yu *et al. Nature* **545**, 224–228; 2017). This process is key to both tumour growth and cardiovascular disease. Wuxi NextCODE uses AI as part of its approach of classifying genes according to their roles and other attributes, to look for connections between RNA-sequence variations, expression levels, molecular function and gene location. Using this approach, Chittenden’s team discovered that FGFs exert their influence through the control of glucose metabolism.

Some think the potential of AI to pinpoint previously unknown causes of disease will accelerate the trend towards treatments designed for patients with specific biological profiles. “Personalized medicine has been talked about for a long time,” says Hunter. “AI is going to enable it.”

Sceptics point out that some of these more enthusiastic claims echo the excitement over computer-aided drug design, which began in the early 1980s. Although such *in silico* modelling techniques are important in modern drug research and development (R&D), they have not halted a decline in pharmaceutical-industry R&D productivity dating back to the mid-1990s.

MOVING GOALPOSTS

Whatever happens, industry leaders agree that drug-discovery jobs and the skills needed to do them are unlikely to remain the same. Some think that broader training is needed. Narain says that “there needs to be a radical shift” in the way PhDs and other graduate courses are conducted, and that this should extend to medical-school and undergraduate teaching. He adds, “The years of students focusing solely on — and learning more than anyone else about — a particular gene mutation, say, are over.” Chittenden agrees: “The PhD is going to look very different ten years from now. Academic curricula will be broader. The next generation needs, first and foremost, the understanding of human biology, but coupled with computer science, computational statistics and statistical machine learning.”

Others think it is more a case of picking up the basics without diverting attention from core areas of expertise. “Undergraduates in biology need to move towards basic competency in statistics and computational ideas,” says Russ Altman, a biomedical AI researcher at Stanford University in California. “But at PhD level, people need to acquire deep, technical skills. They will be paid for depth, not breadth.”

In 2003, Altman co-launched an undergraduate degree in biomedical computation for students who want to delve deeply into both disciplines. It was relaunched within his institution’s bioengineering department in March. “I think that at Stanford we’re getting an early look at what is going to be happening at campuses worldwide,” he says.

There is little consensus about how, even

HERE COME THE ROBOTS

A new type of scientist

When the time comes for the history of artificial intelligence (AI) to be written, the algorithm that gets the job is likely to flag 12 June 2007 as worthy of note. That was the day that a robot called Adam ended humanity's monopoly on the discovery of scientific knowledge — by identifying the function of a yeast gene.

By searching public databases, Adam generated hypotheses about which genes code for key enzymes that catalyse reactions in the yeast *Saccharomyces cerevisiae*, and used robotics to physically test its predictions in a lab. Researchers at the UK universities of Aberystwyth and Cambridge then independently tested Adam's hypotheses about the functions of 19 genes; 9 were new and accurate, and only 1 was wrong.

"Robot scientists using AI can test more compounds, and do so with improved accuracy and reproducibility, and exhaustive, searchable record-keeping," says systems biologist Steve Oliver of the University of Cambridge, a member of the group that developed Adam.

In January, the same team announced that Adam's more advanced robot colleague, Eve, had discovered that triclosan, a common ingredient in toothpaste, could potentially treat drug-resistant malaria parasites. The researchers developed strains of yeast in which genes essential for growth had been replaced with their equivalents either from malaria parasites or from humans. Eve then screened thousands of compounds to find those that halted or severely slowed the growth of the strains dependent on the malaria genes but not those containing the human genes — to target the parasites while reducing the risk of toxicity. Early results were used to inform the selection of later candidates to screen.

This identified triclosan as affecting malaria-parasite growth by inhibiting the DHFR enzyme — also the target of the antimalarial drug pyrimethamine. However, resistance to pyrimethamine is common. The researchers showed that triclosan could act on DHFR even in pyrimethamine-resistant parasites. **N.F.**

just a decade from now, AI will affect the skills needed to discover the therapies of the future. "Being able to code will be useful for at least the next 5–10 years, but my suspicion is that, beyond that, computers will largely do it for us," says computational medicinal chemist Anthony Bradley at the University of Oxford, UK. "In the lab, we might need a more highly trained, specialized workforce working with the automation and AI experts to fine-tune processes in particular reaction areas," he says. Or, he adds, it might be that wet-lab skills (those needed to perform practical chemical or biological experiments) "might be no use ten years from now".

Bradley uses the Diamond Light Source synchrotron near Oxford to screen compounds for small chemical fragments that bind to molecular targets, even if only weakly, with the aim of improving their binding strength to produce new therapies. He is a member of a group that is using artificial neural networks — an approach to training algorithms inspired by the way our brains process information — as part of a structure-based drug-design project with the Oxford Protein Informatics Group. The aim is to use publicly available data on the structural and chemical activity of small molecules to teach their system to identify those that will act on protein drug targets.

What can those hoping to work in drug discovery do to prepare themselves for this rapidly evolving environment? Taking steps to become informed and flexible are important, say those at the cutting edge of the field.

"My training gave me the groundwork so that I knew roughly where the field was, but to some extent it's down to students themselves to see the way technological trends are going," says Bradley. "Only by remaining versatile can you make the best use of the power of the available tools." He advises those seeking to enter the drug-discovery field to keep track of developments in AI by monitoring the latest articles in leading journals and technology-focused news sources and blogs.

Self-driven learning is especially important, Bradley says, because there are limits to how well universities can provide the skills that students need to be ready for the future role of AI in research. "Almost by definition," he says, "no one can really know what those skills will be."

Some of the more extravagant predictions being made about the ability of AI to revolutionize drug discovery might well turn out to be overblown. Critics point out that there are commercial interests at play, and that, as yet, there are no approved AI-developed drugs. Narain, who thinks the technology will drive major advances, agrees that overblown claims are being made, but says it won't be long before these are exposed for what they are. "The hype can't last very long because over the next five years or so, the truth will come out in the data," he says. "If by then we are creating better drugs, and doing it faster and cheaper, then AI will really take off." ■

Nic Fleming is a freelance science writer based in Bristol, UK.